OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 66, 5 (2004) 0305-9049

Truth and Robustness in Cross-country Growth Regressions*

KEVIN D. HOOVER† and STEPHEN J. PEREZ‡

†*Department of Economics, University of California, Davis, CA, USA*
(e-mail: kdhoover@ucdavis.edu)
‡*Department of Economics, California State University, Sacramento, CA, USA*
(e-mail: sjperez@csus.edu)

Abstract

We re-examine studies of cross-country growth regressions by Levine and Renelt (*American Economic Review*, Vol. 82, 1992, pp. 942–963) and Sala-i-Martin (*American Economic Review*, Vol. 87, 1997a, pp. 178–183; Economics Department, Columbia, University, 1997b). In a realistic Monte Carlo experiment, their variants of Edward Leamer's extreme-bounds analysis are compared with a cross-sectional version of the general-to-specific search methodology associated with the LSE approach to econometrics. Levine and Renelt's method has low size and low power, while Sala-i-Martin's method has high size and high power. The general-to-specific methodology is shown to have a near nominal size and high power. Sala-i-Martin's method and the general-to-specific method are then applied to the actual data from Sala-i-Martin's original study.

I. Growth regressions and the problem of robustness

Economists typically prefer theoretically informed empirical investigations. Sometimes, however, we face questions for which there is no generally

*For helpful comments on earlier drafts, we thank Oscar Jorda, Judith Giles, Wayne Joerding, and participants in seminars at the University of California, Irvine, the University of Victoria, the University of Oregon, and the University of Hawaii, Jonathan Temple, David Hendry, Katarina Juselius, Søren Johanssen, the organizers and participants in the conference, 'Bridging Economics and Econometrics: Empirical Applications and Econometric Methods' at the European University Institute, 6–9 June 2001, and an anonymous referee. We also thank Orley Ashenfelter for his help in getting this project off the ground, and Jeannine Henderson and Michael Dowell for able research assistance.

JEL Classification numbers: C4, C8, O4.

agreed upon constrained optimization model to which the empirical researcher can turn. Two, often related, responses to this situation are common. First, economists sometimes take a broader view of the 'theory' that they aim to test, to include less formal considerations (e.g. factors drawn from political science or sociology). Secondly, they sometimes take what, from the point of view traditional econometrics, is an atheoretical approach. Both responses are well exemplified in a series of empirical investigations that are referred to as 'cross-country growth regressions'. In this literature, cross-sectional regression or panel-data techniques are used to identify which of a large number of factors are statistically and economically significant determinants of growth rates.[1]

One problem with the literature is that different studies reach different conclusions depending on what combination of regressors the investigator chooses to put into his regression. In an attempt to put some order into the literature, Levine and Renelt (1992) assembled a cross-sectional data set with a large number of potential regressors and subjected it to a variant of Leamer's (1983, 1985) 'extreme-bounds analysis'.[2] Subsequently, Sala-i-Martin (1997a, b) criticized Levine and Renelt's method and suggested his own, less restrictive variant on extreme-bounds analysis.

Our investigation has two goals. First, we compare the effectiveness in a realistic Monte Carlo simulation of each of these extreme-bounds methodologies to that of a mechanized version of the general-to-specific specification search methodology associated with David Hendry and others and often referred to as the London School of Economics (LSE) methodology.[3] We then apply the LSE approach and Sala-i-Martin's (1997a, b) variant of the extreme-bounds methodology to the specification of cross-country growth regressions.

II. Alternative search methodologies

We begin by describing the competing search methodologies.

[1]The literature is huge. Important contributions are due to Kormendi and Meguire (1985), Grier and Tullock (1989), Barro (1991), DeLong and Summers (1991) and Sachs and Warner (1995, 1996). A recent book by Barro (1997) gives a good overview to the literature.

[2]Temple (2000) re-examines Levine and Renelt's (1992) data set from a perspective largely sympathetic to extreme-bounds analysis, including Sala-i-Martin's variant.

[3]The adjective 'LSE' is, to some extent, a misnomer. It is derived from the fact that there is a tradition of time-series econometrics that began in the 1960s at the London School of Economics (see Mizon, 1995, for a brief history). The practitioners of LSE econometrics are now widely dispersed among academic institutions throughout Britain and the world. The LSE approach is described sympathetically in Gilbert (1986), Hendry (1987, 1995, especially chapters 9–15), Pagan (1987), Phillips (1988), and Ericsson, Campos and Tran (1990). For more sceptical accounts, see Hansen (1996), Faust and Whiteman (1995, 1997) to which Hendry (1997) replies.

Two variants on extreme bounds

The central idea in Leamer's (1983, 1985) analysis is that a coefficient of interest is robust only to the degree that it displays a small variation to the presence or absence of other regressors. Leamer and Leonard (1983) define the extreme-bounds for the coefficient of a particular variable within a search universe as ranging between the lowest estimate of its value minus two times its standard error to the highest estimate of its value plus two times its standard error, where the extreme values are drawn from the set of every possible subset of regressors that include the variable of interest. A variable is said to be *robust* if its extreme bounds lie strictly to one side or the other of zero. And the narrower the extreme bounds, the more confidence one is supposed to have in the coefficient estimate.

Employing a modified version of Leamer's (1983, 1985) approach that reduces the number of regressions needed to compute the extreme bounds, Levine and Renelt (1992) find that few variables can be regarded as robust determinants of economic growth (i.e. almost all coefficient estimates are 'fragile').

Sala-i-Martin (1997a, b) argues that Levine and Renelt (1992) employ too strict a standard of robustness. He suggests that if *most* of the distribution of the coefficient estimates (±2 SE) lie to one side of zero, then the rest might be regarded as irrelevant outliers, and the variable should be regarded to be robust. By analogy with the ordinary practice with significance tests, he suggests that we regard a variable to be robust if 95% of the distribution of the coefficient estimates lies to one or the other side of zero. Predictably, on this more permissive standard, Sala-i-Martin (1997a, b) finds considerably more variables to be robust determinants of economic growth.

Leamer's (1983, 1985) notion of robustness strikes us as an odd one. There is no reason to believe that a variable that is robust in Leamer's (1983, 1985) sense is thereby guaranteed to be a true determinant of economic growth or that a true determinant of economic growth is guaranteed to be a robust one, or even that there is a high correspondence of any kind between truth and robustness (see Hoover, 1995, Hoover and Perez (2000).[4] Leamer (in Hendry *et al.*, 1990, p. 188) rejects the notion of a true specification: 'I … don't think there is a true *data-generating process* …,'. But then it is puzzling what one is supposed to do with one's robust coefficient estimates.

A practical question to ask of any search methodology, such as Levine and Renelt's (1992) or Sala-i-Martin's (1997a, b) versions of extreme-bounds

[4]Closely related criticisms of extreme-bounds analysis are made by Breusch (1990) and Hendry and Mizon (1990).

analysis, is: when there is a truth to be found, does the methodology discover it? Advocates of extreme-bounds analysis sometimes argue that the concern for truth is misplaced, because to reject a variable as fragile is not to deny that it might be a true determinant of the dependent variable. Rather it is to deny that we have good evidence for its truth; i.e. the claim is about us and our certainty, not about the world. But this looks at the wrong side of the issue. The extreme-bounds procedure also identifies some variables as robust. If this says that our evidence is good, that we are more certain, we are still entitled to ask: Of what is our evidence good evidence? What are we more certain about? Surely, the evidence is evidence about whether or not some variables are true determinants of the dependent variable, and the degree of certainty is a measure of how much stock we should place in the truth of the specification.

The required notion of truth is not a metaphysical puzzle. A variable is a true determinant of economic growth if variations in that variable (induced by policy or accident) can be relied upon to yield predictable variations in the rate of economic growth. To discover such determinants we seek to convince ourselves that particular variables predictably explain past growth rates; and then we hope that the relationship is an enduring one that can be used to explain future growth rates. Whether we are warranted in claiming that particular determinants are true or not is a nice question for the philosophy of science perhaps. Nevertheless, when we use estimated relationships instrumentally, we must be assuming that they are true in this sense and not just correlations or data summaries.

Although, for the reasons just stated, we disagree with their arguments, we recognize that serious econometricians maintain that robustness, while not a measure of truth, is nevertheless a property which conveys useful information about the uncertainty surrounding model specification. Temple (2000) provides a good example of this argument. He maintains that extreme-bounds analysis appropriately modified to account for parameter heterogeneity, model uncertainty, and outliers 'is a useful way of communicating any uncertainty surrounding the choice of the model, and hence uncertainty surrounding parameter estimates and standard errors' (Temple 2000, p. 201).

McAleer (1994, p. 347, 349) and Granger and Uhlig (1990) attempt to sharpen the information available in model comparisons by eliminating from the comparison set models that are clearly poor (as judged, for example, by low R^2). Brock and Durlauf (2001, p. 235, especially Footnote 3) accept the role of extreme bounds in indicating specification uncertainty, but argue that it is an inefficient method of specification search because of collinearity. They appeal (p. 262) to model averaging techniques as a means of more effectively assessing and *reporting* the robustness of regression results. Bayesian model

averaging is developed *inter alia* by Raftery, Madigan and Hoeting (1997) and applied to issues related to growth by Doppelhofer, Miller and Sala-i-Martin (2000) and by Fernandez, Ley and Steel (2001).

While we maintain a firm, negative view about the methodological utility of extreme-bounds analysis (see especially, Hoover and Perez, 2000) as a means of specification search, we recognize that others find it valuable as a means of communicating model uncertainty. In the end, we believe that it is unnecessary to resolve the matter here. Instead, we address the narrower question of the effectiveness of extreme-bounds methods as a means of identifying the true data-generating process.

General-to-specific

In the linear context typical of the cross-country growth literature, the general-to-specific methodology begins with the idea that the truth can be characterized by a sufficiently rich regression: the *general regression*. In particular, if every possible variable is included in the regression, then the regression must contain all the information about the true determinants. It may, however, not provide it in a perspicacious form. The information content might be sharpened by a more parsimonious regression – the *specific regression*. This specific regression is acceptable if it has the properties (a) it is statistically well specified (for example, it has white noise errors); (b) that it is a valid restriction of the general regression, and (c) that it encompasses every other parsimonious regression that is a valid restriction of the general regression.[5]

One regression *encompasses* another if it contains all the information of the other regression.[6] LSE econometricians have developed various ways of implementing encompassing tests, but there is an easy way to understand the general notion (and a simple way to implement a test). Consider two competing models of the same dependent variable. If we form a third model which uses the union (excluding redundant variables) of both sets of regressors, then each original regression can be seen as nested in the joint regression. If one of the original regressions can be shown to be a valid restriction of the joint regression, then it encompasses the other regression. Of course, it may turn out that neither regression encompasses the other. There may be some third regression, more parsimonious than the joint regression

[5]The general-to-specific methodology is explained in detail by *inter alia* Phillips (1988) and Hendry (1995). Bleaney and Nishiyama (2002) apply the notion of encompassing to growth regressions.

[6]For general discussions of encompassing, see, for example, Mizon (1984), Mizon and Richard (1986), Hendry and Richard (1987), Hendry (1988, 1995, chapter 14).

that encompasses them both or it may be that the joint regression is as parsimonious as the joint set of regressors permits.

The LSE approach has been used almost exclusively in time-series contexts. There is, however, nothing in its conceptual structure that prevents its extension to cross-sectional data.

Previously, we have evaluated the efficacy of the general-to-specific approach (Hoover and Perez 1999).[7] Our goal was to determine whether common objections to the LSE approach had practical merit. One objection is that any path of simplifications from the general to the specific is just one of many and there is no guarantee that a particular simplification will be the true specification. We acknowledge the problem, but we showed that simple methods of generating a feasible number of competing specifications and then choosing among them on the basis of encompassing tests was an effective strategy.[8]

A second objection is that the general-to-specific searches involve multiple testing with unknown distributional properties. In particular, many conjecture that the size of the whole search procedure is vastly larger than the conventional sizes of the underlying specification tests. This is the objection that is usually associated with a general condemnation of 'data mining'. The objection is often framed as in Lovell (1983) in the context of simpler search procedures (for example, stepwise regression, max R^2 maximin t-statistics) that lack key features of the LSE general-to-specific approach – particularly attention to encompassing the general model and to rigorous specification testing (see Hoover and Perez, 1999, 2000, for further discussion). In our study, we found in a realistic Monte Carlo setting that the size of the general-to-specific search was very close to the nominal size of the underlying specification tests. This suggests that a distinction must be drawn between undisciplined or wrongly disciplined data mining, which is invidious, and well disciplined data mining, which is useful in many contexts.

The LSE approach is not a mechanical one. Instead it relies on a combination of system and econometrician's art. Our evaluation was necessarily based on a mechanical approximation to what LSE econometricians actually do. Some aspects of the approach that might make it more successful were ignored. Nevertheless, the overall assessment was positive.

[7]Our article was the subject of a symposium in the *Econometrics Journal* (1999: no. 2), which included a comment by Hendry and Krolzig (1999). Subsequently, Krolzig and Hendry (2001) refined our algorithms and provided further evidence of their efficacy in the time series context. Their refinements are embodied in a program, PcGets (Hendry and Krolzig 2001). This study was largely completed before PcGets became available.

[8]We regarded multiple search paths as an innovation relative to the LSE approach. In response to our suggestion, Krolzig and Hendry (2001) adopt our approach and Hendry and Krolzig (1999) point out a precedent (Mizon 1977).

III. The effectiveness of three search methodologies in a realistic Monte Carlo study

In this section, we adapt the LSE approach (described in greater detail below) to a cross-sectional context. Our goal is to evaluate the success of two versions of extreme-bounds analysis in a realistic Monte Carlo setting in which we know in fact what the true determinants are. The setting is 'realistic' in the sense that it uses actual variables, rather than ones fabricated from random-number generators, as the true determinants and the other variables in the search universe. The dependent variable, which is calibrated to act like the rate of economic growth, is generated through a bootstrap procedure. The realistic setting ensures that the problem faced in the Monte Carlo is similar to the one that an actual investigator of the determinants of economic growth faces.

We begin, starting in the next subsection, with a description of the data to be used in the simulations and, after that, of the simulations themselves. We then proceed to a description of the detailed implementation of the two variants of extreme-bounds analysis and of the general-to-specific algorithm. Finally, we present the results of the simulation study.

Data and simulated data

In order to understand the effectiveness of search methodologies, it is essential that the variables of the search universe used in the simulations display the same sort of intercorrelations as actual data. To achieve this we start with the data set used by Levine and Renelt (1992). Their original data set contains 40 variables for 119 countries. Reporting is not complete so that the 119×40 matrix of variables has many missing values. A number of variables and countries are, therefore, deleted from the data set with the aim of producing the largest complete matrix possible. The result is a 107 country by 36 variable data set. The original Levine-and-Renelt data set and the reduced, complete data set are described in detail and are downloadable from our websites (http://www.econ.ucdavis.edu/faculty/kdhoover/research.html and http://www.csus.edu/indiv/p/perezs/Data/data.htm). The average rate of growth of GDP per capita for 1960–89 is the variable of interest in the simulations. Each simulation replaces actual GDP growth with a simulated variable constructed from a linear combination of a subset of the other variables in the data set.

Alternative specifications and the criteria of success

In earlier work on time-series models, Hoover and Perez (1999) used the particular set of models suggested by Lovell (1983) as the 'true' specifications.

The particular models were suggested by competing theoretical approaches in macroeconomics and reflected several possible dynamic specifications. That approach is less natural in this context, because the literature is motivated largely by *a priori* ignorance of the empirical factors that might explain growth.

One issue that arose in the time-series context does carry over. Any search procedure may fail to select a variable for one of three reasons: (1) it, in fact, is not a true determinant of the dependent variable or (2) the search method is unsuccessful or (3) the signal-to-noise ratio is low (i.e. there is insufficient variance in the independent variable relative to the dependent variable). The first type of failure is desirable, the second clearly not. The third, however, is unavoidable. As there is no reason to believe that a true specification would necessarily have only those variables that are easy to detect, there is no reason in evaluating different search procedures to favour specifications in which all the variables have a high signal-to-noise ratio. The failure to identify a variable with a low signal-to-noise ratio should not be regarded as a failure of the method. Of course, this is a matter of degree. The strategy we adopt is to simulate models with randomly chosen specifications and to evaluate their success relative to norms that depend on the signal-to-noise ratios for each independent variable.

Each simulation is based on a true specification that relates the growth rate to zero, three, seven, or 14 independent variables. Let the growth rate be $\mathbf{y}$ (a 107×1 vector). Let the complete set of variables be in the data set $\mathbf{X}$ (an 107×34 matrix). Let $\mathbf{X}_j$ be a randomly selected j-element subset of the variables of $\mathbf{X}$, where $j = 0, 3, 7, 14$. There is a degree of unavoidable arbitrariness in the choice of values for j. We justify our choices as follows: zero independent variables is a baseline for checking the size of the search. Levine and Renelt (1992) and Sala-i-Martin (1997a, b) consider regressions with no fewer than three and no more than seven independent variables. We, therefore, study simulations in which there are three and seven variables as well. When the simulated true specification has either three or seven variables, one of Levine and Renelt's or Sala-i-Martin's test specifications can, in fact, coincide precisely with the truth. Levine and Renelt's and Sala-i-Martin's search algorithms do not prevent *more than seven* variables from being identified as robust. In fact, Sala-i-Martin (1997a) identifies 21 variables as robust (and maintains an additional three as free variables in all regressions, for a total of 24 selected determinants of country growth rates). True specifications with 14 variables (chosen because it is twice seven), therefore, allow for the implied cases in which the test specifications are, to different degrees, underspecified relative to the truth.

For each j, 30 different *specifications* are chosen. And for each specification, 100 *simulations* are run.[9] We proceed as follows:

1. Select a j-element subset. This defines the specification.
2. Run the regression, $\mathbf{y} = \mathbf{X}_j\mathbf{B} + \mathbf{u}$ and retain the estimates of the coefficient matrix, $\hat{\mathbf{B}}$, and the estimated residuals, $\hat{\mathbf{u}}$.
3. For each simulation search $i = 1, 2, \ldots, 100$ construct a simulated dependent variable $\mathbf{y}_i^* = \mathbf{X}_j\hat{\mathbf{B}} + \hat{\mathbf{u}}_i^*$. As the Monte Carlo is based on actual data, which may be heteroscedastic, we construct the elements of the vector $\hat{\mathbf{u}}_i^*$ by sampling from $\hat{\mathbf{u}}$ using a wild bootstrap.[10]
4. The three search procedures are run for the ith simulation. The successes and failures at identifying the true variables are recorded for each search procedure.
5. The procedure begins again with a new simulation at step 2 until 100 simulations are recorded. The type I and type II errors are recorded.[11]

The proportion of type I error, the *empirical size*, is calculated as the ratio of the incorrect variables included (significantly at the 5% critical level for general-to-specific searches) to the total possible incorrect variables. The empirical size for a given specification measures what proportion of variables with true coefficients equal to zero are chosen in the specification. We report the *size ratio*, defined as the ratio of the empirical size to the nominal size of the exclusion tests used in the search algorithm (0.05 in all the results we report). A size ratio of unity implies that the size of the search procedure is exactly the nominal size.

The *empirical power* for a given true variable is the fraction of the replications in which the variable is picked out by the search procedure (significantly at the 5% critical level for the general-to-specific procedure); i.e. it is the complement of the proportion of type II error. In order to control for variations in the signal-to-noise ratio, we compute the *true* (*simulated*) *power* from the proportion of type II error for each specification over the 100 bootstrap simulations without search (i.e. with knowledge of the correct regressors). The true (simulated) power for a given true variable is the empirical power that one would estimate if there were no uncertainty about the

[9]Simulations use Matlab 5.2 running on PC with 300 Mz. We would have preferred to examine both more specifications and more simulations of each specification. Unfortunately, each run of 100 simulations for one specification using all three search procedures takes about one and a half days of computing time.

[10]Our implementation follows Brownstone and Kazimi (1998, section 2). The wild bootstrap is due to Wu (1986). Horowitz (1997) shows that the wild bootstrap is superior to the more familiar paired bootstrap when data are heteroscedastic.

[11]We recognize that the object of our investigations is not a classical test statistic of the kind to which the terms 'size' and 'power' normally attach. However, we believe that the statistics that we report based on measures of the type I and type II error are so clearly analogous to size and power that it is natural to use these terms in this context.

true specification, but sampling uncertainty remained. If the signal-to-noise ratio is low, the true (simulated) power will also be low; and, if it is high, the true (simulated) power will be high. The *power ratio* is defined as (*empirical power*)/*true*(*simulated*)*power*. A power ratio of unity indicates that a search algorithm does as well at picking out the true variables as one would do, given the signal-to-noise ratio, with full knowledge of the true specification.

There is, of course, always a balancing between type I and type II error. If one put no weight on type I error, a search algorithm can achieve 100% power by selecting every variable in the data set. The power ratio in that case could easily be much greater than unity, as the true (simulated) power is sometimes very low. The cost, of course, is that the size of such an algorithm is large. Similarly, if one puts no weight on type II error, a search algorithm can achieve a low size by selecting nothing. The cost is that the power of such an algorithm is zero.

Extreme-bounds analysis

We assess particular variants of Leamer's extreme-bounds analysis related to those used by Levine and Renelt (1992) and Sala-i-Martin (1997a, b). A practical problem in implementing extreme-bounds analysis is the large number of regressors. For example, Levine and Renelt's (1992) data set has 39 variables excluding the dependent variable. There are $2^{39} = 5.498 \times 10^{11}$ linear combinations of the regressors. At one second per regression, it would take 17,433 years to try them all.

Levine and Renelt simplify the problem by adopting Leamer's notion that some variables should be included in every regression on the assumption that they are known to be robust *a priori*. The variables real per capita GDP in 1960, primary school enrollment rate in 1960, and the average investment share of GDP 1960–1989 are included in every regression. In describing Leamer's approach, McAleer, Pagan and Volcker (1985) divide the universe of regressors into *free variables*, which theory dictates should be in the regression; *focus variables*, a subset of free variables that are of immediate interest; and *doubtful variables*, which competing theories suggest might be important. Levine and Renelt treat the three variables included in every regression as free variables, and they let every other variable in turn play the part of a focus variable, while linear combinations of the remaining variables play the part of doubtful variables. They restrict the number of subsets of the doubtful variables further by considering only subsets with three or fewer variables. The largest regression for Levine and Renelt, then, has seven independent variables, exclusive of the constant term: one focus variable, three free variables, and (at most) three doubtful variables.

Sala-i-Martin's (1997a, b) approach modifies Levine and Renelt's search procedure in two ways. First, he considers only regressions of *exactly* seven independent variables: one focus variable, three free variables, and (exactly) three doubtful variables. He tries every linear combination of three doubtful variables in the search universe. Secondly, he looks at a different criterion for robustness. The estimate of a coefficient on a focus variable is robust in Sala-i-Martin's sense if 95% or more of the estimates (±2 SEs) lie to one side of zero.

The results reported here follow Sala-i-Martin's evaluation procedure modified to eliminate the free variables. To compute the extreme bounds each variable in the search universe is allowed to be the focus variable in turn and regressions that include it and every subset of exactly three other variables (plus a constant) is computed. From these estimates, variables are identified as robust on the Levine and Renelt and the Sala-i-Martin criteria.

Our procedure differs from both Levine and Renelt, because we do not maintain that we have a priori knowledge of *any* of the true regressors. This seems reasonable in the simulations as the true specifications are chosen randomly. We did, however, examine another set of simulations (not reported in detail here) in which the three free variables are part of every true specification and are maintained in every search. There is no qualitative difference between these simulations and the ones reported here.

General-to-specific

The precise details of the general-to-specific algorithm are given in Appendix A. Here we provide an outline of the procedure. The search procedure proposed is a modification for the cross-sectional context of the general-to-specific search procedure evaluated in Hoover and Perez (1999) in a time series context.

There are five principal elements in the search procedure.

First, the data are divided randomly into two overlapping samples, each with 90% of the data. A search is conducted over each subsample and only those variables that are selected in both subsamples are part of the final specification.[12]

Secondly, each search begins with a general specification in which all the variables in the search universe are included as regressors. The general specification is simplified sequentially by removing variables with low

[12]This procedure was an innovation in Hoover and Perez (1999) relative to the LSE methodology, but has been adopted in, for example, Hendry and Krolzig (1999, 2001), and Krolzig and Hendry (2001). In a study that has only just become available to us, Hendry and Krolzig (2003) conclude that full-sample search dominates the subsample procedure when both are conducted at the same size. The procedure does not improve the size/power tradeoff although it is successful at controlling the size for selection problems.

t-statistics one at a time. Initially, five simplification paths are tried in which each of the variables with the five lowest *t*-statistics is the first variable to be removed along a simplification path. After that – with the exceptions noted below – variables with the lowest *t*-statistic are removed one at a time until all the remaining variables are significant on a 5% test. After removal of each variable, a battery of specification tests is performed. The test battery includes a Breusch and Pagan (1980) test for heteroscedasticity, a subsample stability test using an equality of variance test (a cross-sectional analogue to a Chow test), and an *F*-test of the restrictions from the general model. The number of tests failed is recorded for each step.

Thirdly, after all variables in a specification are significant, the test battery is run. If all tests are passed, this is the terminal specification. If any are failed, the last specification passing all tests becomes the current specification.

Fourthly, a new round of variable elimination proceeds with the removal of the variable with the lowest *t*-statistic in the chosen specification. At each step the test battery is run. If a specification fails one of the tests, the last removed variable is replaced, the variable with the next lowest *t*-statistic is removed and the test battery is run again. This process continues until a variable can be removed without failing any of the tests or all variables are tried.

Fifthly, once all search paths have ended in a terminal specification, the *final specification* is chosen through a sequence of encompassing tests. We form the non-redundant joint model from each of the different terminal specifications; take all candidate specifications and perform the *F*-test for encompassing the other specifications. If only one specification passes, it is the final specification. If more than one specification passes, the specification with the minimum Schwarz criterion is the final specification. If no model passes, reopen the search on the non-redundant joint model (including testing against the general specification) using only a single search path and take the resulting model as the terminal specification. The *final specification* is, as noted above, the intersection of the regressors of the overall terminal specifications from the two 90% subsamples.

Results of the simulations

The results of the simulations are presented in Table 1. Recall that for both the size and the power ratio a value of unity is a useful reference point. A size ratio of unity indicates that the algorithm incorrectly accepts a variable at the same rate that independent tests of a 5% nominal size would do. A power ratio of unity indicates that the algorithm chooses the true variables at the same frequency that one would if one knew the true specification and used a *t*-test with a 5% critical value to decide whether a variable should be retained.

TABLE 1

The efficacy of three search algorithms

	Extreme-bounds analysis		*Modified extreme-bounds analysis*		*General-to-specific*	
Models with:	*Size ratio**	*Power ratio†*	*Size ratio**	*Power ratio†*	*Size ratio**	*Power ratio†*
0 true variables	0.060		1.10		0.75	
3 true variables	0.003	0.43	5.17	0.77	0.77	0.95
7 true variables	0.030	0.13	5.89	1.10	0.81	0.93
14 true variables	0.020	0.04	5.45	0.67	1.02	0.82

Notes:

The basic data are a pool of 34 variables described in Memorandum 1 downloadable from our websites (http://www.econ.ucdavis.edu/faculty/kdhoover/research.html and http://www.csus.edu/indiv/p/perezs/Data/data.htm). For each number of true variables, 30 models are specified by choosing the indicated number of regressors at random from the pool. Coefficients are calibrated from a regression of the chosen regressors on the actual average growth rate. One hundred dependent variables are created from the same regressors and coefficients and error terms constructed with a wild bootstrap procedure from the errors of the calibrating regression. Specification searches are then conducted by each of the three methods and the number of type I and type II errors are recorded. Statistics reported here average over each of the 100 simulations for each of the 30 models. Details of the simulations and the search procedures are found in section 2 and Appendix A.

*Size is calculated as the proportion of incorrect variables included (significantly for general-to-specific) to the total possible incorrect variables. The size ratio is average ratio of the size to the nominal size (0.05) used as the critical value in all the hypothesis tests in the search procedures. A size ratio of 1.00 indicates that on average the size is equal to the nominal size (0.05).

†Power is calculated as the proportion of times a true variables is included (significantly for the general-to-specific procedure). The true (simulated) power is based on the number of type II errors made in 100 simulations of the true model without any search. The power ratio is the average ratio of power to true (simulated) power. A power ratio of 1.00 indicates that on average the power is equal to the true (simulated) power. The power ratio is not relevant when there are no true variables.

The extreme-bounds analysis using Leamer's original criterion shows an extremely low size irrespective of the number of variables in the true specification. This algorithm almost never selects a variable that does not belong. The trade off, however, is that its power ratio is low in all cases, and it too approaches zero as the number of variables in the true specification becomes equal to and then larger than the number of variables in the specifications used to estimate the extreme bounds. While the extreme-bounds algorithm almost never commits type I error, it almost always commits type II error. This confirms generically the criticism made by Sala-i-Martin of Levine and Renelt's use of extreme-bounds analysis. It is overly strict. It says, 'nothing is robust'; and, in so saying, it is unable to find the truth at all.

In contrast, the modified extreme-bounds analysis does almost exactly the reverse. Its size ratio is only 10% greater than the 5% nominal test size when there are no true variables. It rises to an astonishing 440% greater than the nominal test size for specifications with 14 true variables. Compared with the standard extreme-bounds analysis it picks out too many variables that do not

belong to the true specification. Of course, this increases the power ratio. When there are seven variables in the true specification – the same number as in the regressions used to estimate the extreme bounds, the power ratio is 1.10. The algorithm is more likely to pick the true variables than even knowing the true specification would suggest. At three or 14 variables in the true specification, the power ratio falls substantially. Although Sala-i-Martin shows some success in correcting the overly strict character of Levine and Renelt's method, the cure comes at the price of going way too far in the other direction. His method is overly lax. It says, 'many variables are robust' and, in so saying, it is unable to discriminate the true from the false.

The general-to-specific algorithm finds the middle ground. Its size ratio is below unity except when there are 14 variables in the true specification, and then it is only slightly greater at 1.02. Its power ratio is always a little less than unity, but except for the case of three true variables, it is larger than that for the modified extreme-bounds analysis. In comparison with the other two methods, the general-to-specific algorithm not only usually finds the truth nearly as well as one would if God had whispered the true specification in one's ear, but it is also able to discriminate between true and false variables extremely well.

The fact that the empirical size is well behaved (i.e. near unity) for the general-to-specific search algorithm is perhaps the most striking thing about these findings. Many critics of data mining in general, and the general-to-specific methodology in particular, express *a priori* skepticism of the practice of multiple, sequential testing using conventional critical values. Invariably, they predict that such test procedures are bound to understate the true size of the joint test implicit in the search procedure. The evidence here runs in the other direction altogether. Far from the simulations showing that the empirical size is very high, it is, in fact, lower than the nominal test size. These results are broadly consistent with the earlier findings of Hoover and Perez (1999), who found empirical sizes for the general-to-specific algorithm that were greater – but only a little greater – than the nominal sizes of the tests. One way to understand this result is that the disciplines imposed by the various encompassing tests in the search procedure tend to force the final specification to be close to the true specification. And, if one had known the true specification *a priori*, the nominal test sizes would have been correct. Tests based on a specification that is near the true specification have similar size.

IV. Re-examining the data

Although the investigation of the last section used data from a cross-country growth study in order to better mimic real data, it was a true simulation revealing characteristics of the search methodologies and not of the world. In this section, we apply those methodologies – in light of the simulation

study – to the central question of the cross-country growth literature: what explains the differences in growth rates among nations?

Theory has made some headway with this question. The neoclassical growth model (Solow 1956) tells us that in steady state, growth rates depend on the rates of growth of the labour force and of technological progress, yet it gives us little notion of what might determine technological progress, especially when technology must be conceived to include all aspects of social organization that might relate to the effectiveness of production. Out of steady-state, increasing the rate of utilization of factors of production or increasing capital investment can temporarily increase growth rates. Also, the higher the gap between the current level of output and the steady-state level, the higher the growth rate. Models of growth with increasing returns suggest that we cast a wider net, looking at industrial organization, research and development, investment in education and other factors (Romer, 1986, Lucas, 1988 and many others, ably surveyed in Jones, 1998, Barro and Sala-i-Martin, 1995). These models generally assume that some factor is important and try to work out the mechanisms of its influence, but they give little guidance as to which of the many possible factors really influence growth. Substantial room remains for theory to be informed by empirical investigations.

Sala-i-Martin (1997a, b) makes a persuasive case that Levine and Renelt's choice of data introduces endogeneity problems that are inadequately addressed. He assembles a data set less susceptible to those problems. In the next subsection we describe this data set and a multiple imputation procedure (novel in the cross-country growth literature) for dealing with the fact that many variables do not exist for many countries. Next, we apply both the modified extreme-bounds analysis and the general-to-specific search methodology to this data set and evaluate the results in light of the earlier simulation study, asking what conclusions might be drawn about the determinants of growth differentials.

Adapting to real-world data

The previous section cast doubt on the efficacy of extreme-bounds methods in identifying the true determinants of a dependent variable in a case in which those true determinants were in fact known. The general-to-specific methodology did substantially better. What implications would these simulation results have for reasonable conclusions about the actual determinants of cross-country growth differentials? To investigate this question, we apply the general-to-specific methodology to Sala-i-Martin's (1997a, b) data set. We compare the results in each case to those using Sala-i-Martin's modified extreme-bounds analysis.

Sala-i-Martin (1997a, b) correctly observes that Levine and Renelt's data set includes variables that may be endogenous as potential regressors. Endogenous regressors call into question a causal reading, not only of any final regressions based on the data set, but also the validity of the ordinary-least-squares regressions in all the intermediate stages of both search procedures. To account for this in a new data set, Sala-i-Martin, to a greater degree than Levine and Renelt, collected variables that were likely to be predetermined, so that a causal reading of their relationship to the rate of growth of per capita GDP is more plausible.

In the Monte Carlo simulations of the last section, we worked with 'nice' data. In particular, we eliminated a carefully chosen set of countries and variables in order to produce a data matrix without missing values. Sala-i-Martin's data set is missing 14.5% of its values. As Levine and Renelt before him, Sala-i-Martin deals with the missing values through what is sometimes called *casewise* (or *listwise*) *deletion*: for any regression, if a country does not report the values for each of the variables required for that regression, that country is omitted from the regression. Although it is a common practice, casewise deletion presents particular problems in this context.

First, casewise deletion wastes enormous amounts of relevant information. Every country that is missing values for any of the variables in a particular regression is omitted from that regression – and along with it all the data for that country whether they are missing or not. Although only 14.5% of the data is truly missing, Sala-i-Martin's practice of omitting countries with missing data treats a minimum of 25% of the cells as empty and, in the worst case, could treat more than 67% as empty.

Secondly, the millions of regressions of the extreme-bounds analysis are run over a shifting set of countries. The legitimacy of comparing coefficient estimates across regressions conducted on different samples is highly questionable. Yet, extreme-bounds analysis requires a legitimate basis for such comparisons. Similarly, the general-to-specific methodology cannot be implemented properly – even as a mechanical matter – if samples must be constantly shifted, since it vitiates encompassing tests against a general specification.

Our solution to the problem of missing data is to use multiple imputation (see Little and Rubin, 1987; Rubin, 1987, 1996; Schaefer, 1997; and King, Joseph and Scheve, 2001, and the references therein). The details of the procedure are described in Appendix B. The principal advantage of multiple imputation is that it allows us both to effectively fill in the missing data with a maximum likelihood estimate and to retain sampling uncertainty, which helps to support accurate specification tests needed for both extreme-bounds and the general-to-specific methodologies.

Hoover and Perez (1999) show that there is a nonlinear trade-off between power and size as function of the degree of overlap in general-to-specific searches (the greater the overlap, the higher the power, but the higher the size). As the imputation of missing data should have little effect on the size but may lower the power relative to a complete data set (although *not* relative to casewise deletion), we run the general-to-specific searches as described in Appendix B with 90-percent overlap and with 100% overlap (i.e. on a single sample).[13]

Results using the Sala-i-Martin data set

The data set is described in detail and are downloadable from our web-sites (http://www.econ.ucdavis.edu/faculty/kdhoover/research.html and http://www.csus.edu/indiv/p/perezs/Data/data.htm). It contains 64 variables for 138 countries with 14.5% of the values missing. We omit 12 countries for which the data is so sparse that imputation does not seem sensible. The dependent variable is the growth rate of real per capita GDP for 1960–92. We omit the *average age of the population* (mnemonic AGE), because the data do not appear to correspond to its definition: of 138 countries, 107 are reported as 0, 1 as 40, and the remainder as a variety of values greater than 76. And we also omit a variable (mnemonic X0) that duplicates the dependent variable.[14] We are then left with a search universe of 126 countries by 61 variables plus the dependent variable. The search is conducted with the nominal size of all specification tests and *t*-tests set at 5%.

There is no reason to expect that an extreme-bounds analysis of the data set completed using multiple imputation would give the same results as one using casewise deletion. So, for the sake of comparison, Table 2 presents the results of a modified extreme-bounds analysis in a format that corresponds to Table 1 of Sala-i-Martin (1997b), which is an expanded version of Table 1 of Sala-i-Martin (1997a). Running the eye quickly down the columns headed 'Lower Extreme' and 'Upper Extreme' confirms Levine and Renelt's (1992) original conclusion that robustness is rare: only three of the variables are robust on their definition, and these do not include the three free variables. Nevertheless, on the modified robustness criterion – based, following Sala-i-Martin's (1997a) preference, on a non-normal, weighted cumulative distribution function – 13 of

[13]In the simulation studies in section 2, the overlapping samples are chosen randomly. Here the countries omitted from each subsample are chosen by selecting a seed country randomly and then selecting every tenth country. Since the countries are grouped by region, this procedure guarantees that all regions are well represented.

[14]We were able to reproduce Sala-i-Martin's (1997a, b) modified extreme-bounds estimates using the data. AGE was not robust in the modified extreme-bounds analysis and recomputing the modified extreme-bounds analysis omitting it has little effect on the robustness of the remaining variables.

TABLE 2

Extreme bounds analysis and modified extreme bounds analysis of Sala-i-Martin (1997a, b) data

Mnemonic	*Variable name*	*Lower extreme*	*Upper extreme*	*Fraction significant*	*Beta**	*Standard deviation*	*CDF normal***	*CDF non-normal (weighted)†*	*CDF non-normal (not-weighted)††*
Focus variables‡									
YRSOPEN	Number of years open economy	0.0054	0.0380	1.0000	0.0214	0.0052	1.0000	1.0000	1.0000
CONFUC	Fraction Confusist	0.0186	0.1490	1.0000	0.0903	0.0213	1.0000	1.0000	1.0000
BUDDHA	Fraction Buddhist	−0.0093	0.0552	0.9646	0.0277	0.0104	0.9961	0.9943	0.9943
EQINV	Equipment investment	−0.0182	0.3344	0.9823	0.1488	0.0617	0.9920	0.9912	0.9912
PRIEXP70	Primary exports	−0.0399	0.0127	0.9180	−0.0172	0.0071	0.9924	0.9870	0.9871
LAAM	Latin American dummy	−0.0278	0.0134	0.7786	−0.0099	0.0044	0.9881	0.9828	0.9829
PROT	Fraction of Protestant	−0.0403	0.0093	0.4668	−0.0134	0.0067	0.9762	0.9701	0.9701
FRENCH	French colony	−0.0198	0.0070	0.1781	−0.0073	0.0041	0.9607	0.9550	0.9551
SPAIN	Spanish colony	−0.0255	0.0213	0.3548	−0.0085	0.0051	0.9543	0.9351	0.9352
SAFRICA	Sub-Sahara African dummy	−0.0230	0.0078	0.1118	−0.0066	0.0041	0.9453	0.9319	0.9320
DPOP6090	Growth rate of population	−0.0141	0.0044	0.0836	−0.0044	0.0030	0.9316	0.9233	0.9232
REVCOUP	Revolutions and coups	−0.0319	0.0114	0.0180	−0.0099	0.0070	0.9231	0.9174	0.9174
BRIT	British colony	−0.0083	0.0162	0.0498	0.0046	0.0032	0.9268	0.9174	0.9173
GVXDXE52	Public consumption share	−0.0923	0.0381	0.0043	−0.0308	0.0219	0.9200	0.9132	0.9134
RERD	Exchange rate distortions‡‡	−0.2000	0.1000	0.0000	−0.1000	0.0425	0.9047	0.8986	0.8986
CATH	Fraction of Catholic	−0.0311	0.0178	0.0285	−0.0078	0.0058	0.9117	0.8975	0.8977
ENGFRAC	Fraction of population able to speak English	−0.0310	0.0176	0.0428	−0.0081	0.0060	0.9092	0.8967	0.8968
NONEQINV	Non-equipment investment	−0.0622	0.1525	0.0000	0.0418	0.0352	0.8823	0.8785	0.8786
WARDUM	War dummy	−0.0140	0.0070	0.0009	−0.0039	0.0035	0.8687	0.8597	0.8599
RULELAW	Rule of law	−0.0157	0.0336	0.0000	0.0084	0.0077	0.8601	0.8516	0.8519
ECORG	Degree of capitalism	−0.0031	0.0043	0.0000	0.0011	0.0011	0.8544	0.8509	0.8511
MUSLIM	Fraction of Muslim	−0.0197	0.0241	0.0003	0.0063	0.0060	0.8540	0.8455	0.8455
GDE1	Defense spending share	−0.1502	0.2571	0.0000	0.0685	0.0680	0.8431	0.8344	0.8347

continued overleaf

TABLE 2

(continued)

Mnemonic	*Variable name*	*Lower extreme*	*Upper extreme*	*Fraction significant*	*Beta**	*Standard deviation*	*CDF normal***	*CDF non-normal (weighted)†*	*CDF non-normal (not-weighted)††*
PI6089	Average inflation rate§	−0.2000	0.2000	0.0000	−0.0286	0.0314	0.8190	0.8224	0.8225
LLY1	Liquid liabilities to GDP	−0.0148	0.0249	0.0000	0.0065	0.0069	0.8249	0.8182	0.8182
FRAC	Ethnolinguistic fractionalization	−0.0244	0.0140	0.0000	−0.0062	0.0068	0.8175	0.8130	0.8131
ABSLATIT	Absolute latitude	−0.0005	0.0006	0.0032	0.0001	0.0001	0.8219	0.8101	0.8102
FREEOP	Free trade openness	−0.0676	0.1099	0.0000	0.0254	0.0292	0.8081	0.8042	0.8045
STDC6089	Standard deviation of domestic credit§	−0.1000	0.1000	0.0000	−0.209	0.0253	0.7959	0.7943	0.7944
PRIGHTSB	Political rights	−0.0098	0.0049	0.0034	−0.0010	0.0014	0.7692	0.7761	0.7759
OTHFRAC	Fraction of population able to speak a foreign language	−0.0141	0.0199	0.0027	0.0035	0.0044	0.7882	0.7719	0.7720
WORK60L	Ratio of workers to population	−0.0330	0.0242	0.0000	−0.0054	0.0085	0.7391	0.7307	0.7306
STPI6089	Standard deviation of inflation§	−0.0472	0.0001	0.0000	−0.0036	0.0081	0.6739	0.7277	0.7277
AREA	Area (scale effect)§	−0.0031	0.0027	0.0000	−0.0005	0.0008	0.7229	0.7213	0.7213
CIVLIBB	Civil liberties	−0.0064	0.0109	0.0015	−0.0005	0.0015	0.6353	0.7163	0.7158
HINDU	Fraction of Hindu	−0.0407	0.0253	0.0000	−0.0062	0.0114	0.7067	0.7021	0.7023
MINING	Fraction of GDP in mining	−0.0574	0.0592	0.0000	0.0097	0.0188	0.6974	0.7012	0.7010
LFORCE60	Size of labor force (scale effect)¶	−0.0017	0.0019	0.0000	0.0003	0.0006	0.6794	0.7011	0.7011
SCOUT	Outward orientation	−0.0094	0.0115	0.0000	0.0016	0.0033	0.6842	0.6910	0.6913
H60	Higher education enrollment	−0.2036	0.1474	0.0000	−0.0214	0.0527	0.6579	0.6591	0.6588
DEMOC65	Index of democracy	−0.0304	0.0258	0.0000	−0.0038	0.0097	0.6513	0.6524	0.6525
BMS6087	Standard deviation of the black market premium§	−0.0221	0.0149	0.0000	−0.0025	0.0069	0.6433	0.6433	0.6433
S60	Secondary school enrollment	−0.0522	0.0384	0.0000	−0.0042	0.0140	0.6183	0.6322	0.6317
JEW	Fraction of Jewish	−0.0742	0.0715	0.0000	0.0065	0.0211	0.6204	0.6294	0.6292
PYR60	Average years of primary school	−0.0052	0.0050	0.0000	−0.0005	0.0015	0.6200	0.6263	0.6263
GDC6089	Growth of domestic credit§	−0.3000	0.5000	0.0000	0.0354	0.1190	0.6172	0.6145	0.6145
URB60	Urbanization rate	−0.0249	0.0309	0.0000	0.0018	0.0087	0.5806	0.6081	0.6081

continued overleaf

TABLE 2

(continued)

Mnemonic	*Variable name*	*Lower extreme*	*Upper extreme*	*Fraction significant*	*Beta**	*Standard deviation*	*CDF normal***	*CDF non-normal (weighted)†*	*CDF non-normal (not-weighted)††*
SYR60	Average years of secondary school	−0.0112	0.0148	0.0000	0.0009	0.0039	0.5940	0.6081	0.6083
PINSTAB2	Political instability	−0.0398	0.0681	0.0000	0.0043	0.0162	0.6043	0.6034	0.6034
BMP1	Black market premium	−0.0122	0.0182	0.0000	0.0011	0.0048	0.5939	0.5947	0.5943
HYR60	Average years of higher school	−0.0688	0.0780	0.0000	0.0020	0.0216	0.5367	0.5930	0.5928
HUMAN60	Average years of schooling	−0.0037	0.0034	0.0000	−0.0002	0.0010	0.5645	0.5914	0.5912
HUMANYL	Human60 × log(GDP60)§	−0.4000	0.5000	0.0000	−0.0174	0.1000	0.5523	0.5837	0.5835
FREETAR	Tariff restrictions	−0.1953	0.2476	0.0000	0.0038	0.0678	0.5226	0.5815	0.5811
ASSASSP2	Political assassinations	−0.1083	0.1013	0.0000	−0.0025	0.0354	0.5281	0.5672	0.5671
TOT1	Terms of trade	−0.1743	0.1995	0.0000	0.0033	0.0636	0.5206	0.5657	0.5655
GGCFD3	Public investment share	−0.1471	0.1353	0.0000	0.0006	0.0460	0.5049	0.5582	0.5579
GEEREC1	Government education spending share	−0.5060	0.4687	0.0000	−0.0096	0.1751	0.5218	0.5530	0.5530
Free variables‡‡									
P60	Primary school enrollment	−0.0159	0.0415	0.0124	0.0124	0.0090	0.91593	0.91062	0.91076
LIFEE060	Life expectancy	−0.0007	0.0010	0.0000	0.0002	0.0003	0.71442	0.71172	0.71201
GDPSH60L	log(GDP per capita 1960)	−0.8302	0.5187	0.0000	−0.1842	0.2124	0.80708	0.80297	0.80326

Notes:

See Sala-i-Martin (1997a, b) for a general discussion of the methods and the formulae used in this table.

Heavy line divides variables into 'robust' (above) and 'non-robust' (below the line) on a 95% criterion. All regressions include a constant.

*Beta is the estimated weighted average of the coefficient on the focus variable.

**CDF normal is the proportion of the CDF of the estimated coefficient assuming that the distribution of the estimator is normal.

†CDF non-normal (weighted) does not assume normality but weights the estimated CDFs using the integrated likelihood for each regression.

††CDF non-normal (non-weighted) does not assume normality and does not weight the CDFs.

‡'Focus variables' are potentially important contributors to growth.

‡‡'Free variables' are variables included in all regressions.

§Coefficients have been multiplied by 1,000 for ease of reading.

¶Coefficients have been multiplied by 10,000 for ease of reading.

the focus variables and one of the free variables are robust.[15] Only two of these variables are not also found in Sala-i-Martin's (1997a, b) analysis using casewise deletion. But 11 variables found in that earlier analysis are not found in Table 2. This makes sense. Adding countries to an extreme bounds analysis presents many further opportunities for failures of robustness.

Table 3 presents regression equations based on two general-to-specific searches – one using 90% overlapping subsamples and one using the full sample. Except for two additional variables *equipment investment* and *revolutions and coups*, the full-sample search selects the same variables as the overlapping-subsample search. We prefer the full-sample search as the one likely to have the better power. Table 3 also presents a regression that uses the robust focus variables and the free variables (whether robust or not) from the modified extreme bounds search. Comparison of the full-sample and modified extreme-bounds specifications in Table 3 is consistent with the results of the Monte Carlo study summarized in Table 1. Four variables are chosen by both search methodologies; seven additional regressors are chosen only by modified extreme-bounds analysis; one is chosen only by the general-to-specific algorithm. These are patterns of the type one would expect given the size and power ratios for the two search methodologies reported in Table 1.

The full-sample regression in Table 3 is more parsimonious than the regression based on the modified extreme-bounds analysis. The *P*-values for *F*-tests of each regression against a joint regression that includes all 12 variables generated from either search are both 0.05. At a 5% critical value, both are valid restrictions of the joint model and neither encompasses the other. The extreme-bounds regression has a lower standard error, but at the cost of almost double the number of regressors. The Schwarz criterion, which evaluates the trade-off between improvements in fit and loss of parsimony, is lower for the general-to-specific regression. We conclude – although it is close – that the general-to-specific regression is the preferred specification statistically.

To investigate the relationship between the two competing specifications somewhat further, we form the nonredundant union of the full sample general-to-specific model and the extreme-bounds model and then run the general-to-specific search algorithm again. It chooses a model identical to the full sample model in Table 3. This shows that the extreme-bounds specification was not eliminated because of any peculiarity in the search algorithm, but because it failed the criteria of the specification tests in the algorithm.

How important are the various determinants of economic growth economically? In Table 4, we evaluate the contribution of each of the variables selected in the general-to-specific search by multiplying the

[15]The construction of the cumulative distribution function and the weighting scheme are described in Sala-i-Martin (1997a, b).

TABLE 3

Regressions of the growth rate of GDP per capita (GR) using Sala-i-Martin (1997a, b) data

		Specification search method								
		General-to-specific						*Modified extreme bounds*		
		Overlap			*No overlap*					
Variable (mnemonic)	*Name*	*Coefficient*	*t-statistic*	*p-value‡*	*Coefficient*	*t-statistic*	*p-value‡*	*Coefficient*	*t-statistic*	*p-value‡*
	Constant	0.00960	4.95	2.42×10^{-6}	0.0107	3.92	1.43×10^{-4}	0.0373	2.53	0.0130
REVCOUP					−1.24	2.06	0.0410			
CONFUC	Fraction Confucist*	0.0817	3.76	2.52×10^{-4}	0.0769	3.61	4.38×10^{-4}	0.0540	2.54	0.0120
EQINV	Equipment investment (fraction of GDP)*				0.1075	2.53	0.0130	0.0941	2.08	0.0390
PROT	Fraction Protestant*	−0.00770	1.34	0.182	−0.0120	2.17	0.0320	−0.0105	1.81	0.0730
YRSOPEN	Years as an open economy	0.0251	6.05	1.35×10^{-8}	0.0195	4.46	1.72×10^{-5}	0.0179	3.69	3.30×10^{-4}
BUDDHA	Fraction Buddhist*							0.00600	0.600	0.547
FRENCH	French colony (= 1; = 0 otherwise)							−0.00520	1.38	0.169
GDPSH60L	Log(per capita GDP in 1960)							−0.00420	1.77	0.079
LAAM	Latin American country (= 1; = 0 otherwise)							−0.00470	1.15	0.251
LIFEE060	Life expectancy in 1960							4.69×10^{-5}	0.200	0.843
P60	Primary school enrollment in 1960*							0.0107	1.42	0.160
PRIEXP70	Primary export as fraction of exports in 1970*							−0.00820	1.33	0.187

continued overleaf

TABLE 3

(continued)

	Specification search method		
	General-to-specific		
	Overlap	*No overlap*	*Modified extreme bounds*
$\bar{R}^{2*}$†	0.36	0.42	0.46
Standard error of regression*†	1.43	1.36	1.32
Sum of squared residuals*†	2.76	2.45	2.19
Number of observations	138	138	138
Mean of the dependent variable*	1.789	1.789	1.789
Standard deviation of the dependent variable*	1.791	1.791	1.791
Log-likelihood†	392.02	400.24	407.88
Schwarz criterion†	−5.54	−5.59	−5.48

Notes:
Shaded areas identify variables common to both search procedures.
*Original units of the data (natural fractions) converted to percentage points.
†Calculated as the mean of statistic for the five imputed data sets.
‡*p*-value is for a two-sided test.

coefficient for an independent variable by the change in the independent variable between the country in the 25th percentile and the country in the 75th percentile for each variable. This gives some idea of how much each characteristic serves to differentiate higher from lower growth countries.

Our method of displaying the economic magnitudes of the different determinants must be treated with caution. They are not a guide to true counterfactual experiments, as many of the variables in the data set are given by history or geography, and therefore are not malleable, or have dependencies (e.g. shares professing various religions) that restrict the possible values they might take. Variables may also be skewed in a manner that would mislead us about the practical policy implications of moving to extreme values. Instead we interpret Table 4 as descriptive. It provides an accounting for each determinant.

The variables in Table 4 are arranged in ascending order of the effect on the growth rate attributable to each variable. The five variables divide into three groups.

The first group is the single variable indexing the number of revolutions and coups a country faces. Revolutions and coups have a moderate negative effect on growth rates as moving from the 25th percentile to the 75th percentile would decrease a countries growth rate by 0.397 percentage points.

The second group contains two cultural/religious variables. It is a 0.288 percentage point disadvantage to move from the 25th percentile Protestant country to the 75th percentile. The coefficient for fraction Confucist needs to be interpreted with extreme care. In the Sala-i-Martin dataset, there are only six countries with non-zero values: Malaysia, Singapore, Hong Kong, Taiwan, China, and Korea.

There are two political/economic variables. Investment is relatively important.[16] Interestingly, the search eliminates *non-equipment investment* and *public investment* and retains only *equipment investment*. *Equipment investment* is the second highest positive influence. Openness to foreign trade is the most important effect in either direction. An increase of openness from the 25th percentile country (one of 36 countries with no years open) to Australia (the 75th percentile country) which had 68.9 years of openness would increase a country's growth rate 1.378 percentage points.

V. Lessons for methodology; lessons for growth

There are two main points to this study. The first is methodological. Despite the fact that we do not have good *a priori* theory of the determinants of

[16]This confirms the findings of DeLong and Summers (1991) and Temple (1998) among others. Sala-i-Martin (1997b, p. 7) retains investment in his data set despite its endogeneity because it is a central variable in the neoclassical and other growth models. We agree that it its endogeneity renders its interpretation in the final specifications ambiguous.

TABLE 4

The importance of the determinants of growth rates based on the general-to-specific search of the Sala-i-Martin (1997a, b) data

	Characteristics of the data†							
Variable† (mnemonic)	*Name (units)*	*Countries*	*Minimum*	*25th percentile*	*75th percentile*	*Maximum*	*Coefficient†*	*Effect on growth rate from moving from 25th percentile to the 75th percentile*
Dependent variable	*Growth rate of GDP per capita* (percent)*	112	Madagascar −2.07	Rwanda 0.64	Iceland 2.70	Korea 6.62		
REVCOUP	*Coups and revolutions* (number)	133	42 countries 0.0	42 countries 0.0	Honduras 0.32	Bolivia 1.19	−1.240	−0.397
PROT	*Fraction Protestant* (fraction of population)*	135	28 countries 0.0	China 1.00	Zimbabwe 25.0	Iceland 98.0	−0.012	−0.288
CONFUC	*Fraction Confucist* (fraction of population)*	135	129 countries 0.0	129 countries 0.0	129 countries 0.0	Korea 60.0	0.077	0.000
EQINV	*Equipment investment* (fraction of GDP)*	88	Mozambique 0.02	Angola 1.28	Portugal 5.69	Singapore 14.82	0.108	0.476
YRSOPEN	*Number of years as an open economy* (index)*	125	36 countries 0.0	36 countries 0.0	Australia 68.9	10 countries 100.0	0.020	1.378

Notes:

Variables are listed in ascending order of the effect on the growth rate attributable to each variable evaluated at its median value.

*Original units of the data (natural fractions) converted to percentage points.

†Variables and coefficient values are those that were statistically significantly different from zero at the 5% confidence level in the general-to-specific search reported in Table 3.

‡The characteristics of the data are computed using available data for all countries in the Sala-i-Martin (1997a, b) data set.

differences in growth rates between countries, we would like to identify the true determinants. Robustness in Leamer's sense is not an adequate guide to model specification, whatever other uses it might have. Robustness is neither necessary nor sufficient for a regressor to belong to the data-generating process. Extreme-bounds approaches in the form advocated by Levine and Renelt are too stringent and reject the truth too frequently (small size, but low power), while those advocated by Sala-i-Martin are not discriminating and accept the false too frequently along with the true (high power, but large size). In contrast, the general-to-specific specification search methodology is – like Little Bear's bed in the tale of *Goldilocks* – just right: it maintains a size near (and even a little below) the nominal size of the tests used in the search and has power approaching the true power one should find if the specification were not in doubt.

It is sometimes objected that the advantage of the general-to-specific search is illusory because it presupposes (wrongly, it is asserted) that the true specification is nested in the search universe, and that this is unlikely, since the search universe never includes every variable that matters to the dependent variable in any way. This misunderstands both the exercise conducted in this paper and the underlying strategy of the LSE methodology. Of course, the general-to-specific search cannot locate the true specification if the true variables are not available to the search. But equally, there is no reason to suppose that extreme-bounds analysis is any more informative when variables are omitted from its search universe, than when they are included. The argument is that if any of the methods fail to find the truth when it is in fact there to be found, the method is *a fortiori* unsuccessful. If robustness does not correspond to truth when truth is to be had, why should it be regarded as a desirable characteristic when truth is required but unavailable? We can never guarantee that the specifications selected by the general-to-specific approach are true. But the approach is part of a critical, indeed dialectical, methodology. If anyone seriously argues that an important variable has been omitted from the specification, the appropriate response is to add that variable to the search universe and, then, to rerun the search.

It is worth noting that this same critical spirit can be applied to the general-to-specific search algorithm itself. We have presented only a single version of a general approach. While we have shown that it is superior to the two alternatives that we studied, it is not necessarily the best implementation of that approach. We look forward to further refinements and developments – and perhaps to further horse-races against other search methodologies.

The second main conclusion from the study is that in practice extreme-bounds methods are misleading about the determinants of growth. Sala-i-Martin was right to criticize Levine and Renelt (1992) for rejecting too many potential determinants of growth as non-robust. What is more, he is right to

question the exogeneity of a number of the determinants of growth that they consider. However, the evidence of the general-to-specific approach is that his approach selects many variables that probably do not truly determine differences in growth rates. While the modified extreme-bounds analysis selects most of the variables selected by the general-to-specific search, it also selects a set of other variables that add nothing significant to the explanatory power of the specification. Differences in growth rates are more adequately characterized by a much smaller number of variables. The five variables retained by the general-to-specific search correspond reasonably well to *a priori* growth theory and to a reasonable understanding of political and cultural factors.

The general-to-specific search, therefore, reaches more precise conclusions about the determinants of differences in growth rates among countries than does the modified extreme-bounds analysis. There are three messages. First, to some degree cultural characteristics matter. It is a massive advantage to be Confucian and, given the fame of the 'Protestant work ethic', a surprising disadvantage to be Protestant. Unfortunately, such cultural variables are not easily manipulated by public policy. Secondly, a more hopeful message: investment, which can be affected by policy, matters. Thirdly, politics matters: civil disorder is an important antagonist and openness to world trade is an important promoter of economic growth.

What is surprising is how few of the variables matter in the end and how much is left unexplained. The preferred regression explains only 42% of the variability in countries' growth experiences. Although none of the factors identified contradicts common economic understanding of the growth process, what is omitted gives no special support to the most popular classes of growth models. The neoclassical growth model encourages us to expect evidence of conditional convergence – i.e. evidence that *ceteris paribus* the further behind a country was, the faster it would grow (see Mankiw, Romer and Weil, 1992). The search failed to select real per capita GDP in 1960. A negative coefficient on that variable would have provided some evidence of conditional convergence.

In contrast, Sala-i-Martin (1997a, b) did find evidence of conditional convergence – initial income (real per capita GDP in 1960) was a highly robust free variable. However, this result does not hold up when extreme-bounds analysis is applied to the richer data set used in Table 2. As a free variable, initial income appears in every regression, but with only 80% of the parameter estimates on the conditional convergence side of zero, it fails to make Sala-i-Martin's 95% cut-off for robustness. Again, as a free variable, initial income is also a regressor in the modified extreme-bounds specification in Table 3. With a p-value of 0.08, it is significant at the 10%, but not at the 5%, level. When included in the full-sample

genera-to-specific specification, initial income has a *p*-value of 0.22 and the other coefficients are not significantly changed. If the failure to find conditional convergence is theoretically puzzling, it is puzzling for both the general-to-specific and modified extreme-bounds methodologies applied to Sala-i-Martin's data set.

There are other discrepancies between some growth theories and our findings. The new growth models typically suggest that human capital or education variables should increase growth rates. But again, none of these variables was selected. And again, *primary school enrollment* and *life expectancy* are selected by the modified extreme-bounds analysis only because they are retained as free variables by theoretical priors; they do not show up as robust in Table 2 or statistically significant in Table 3.

General-to-specific search methods have proved superior to extreme-bounds methods in isolating the truth – when the truth is to be found. Applied to actual data, they allow us to identify factors important to explaining the differences in growth rates among various countries. Our study, we believe, uses the available cross-sectional data more fully than any previous study. Yet, it also highlights how much more there is to understand.

Final Manuscript Received: January 2004

References

Amemiya, T. (1985). *Advanced Econometrics*, Blackwell, Oxford.

Barro, R. J. (1991). 'Economic growth in a cross-section of countries', *Quarterly Journal of Economics*, Vol. 106, pp. 407–433.

Barro, R. J. (1997). *Determinants of Economic Growth*, MIT Press, Cambridge, MA.

Barro, R. J. and Sala-i-Martin, X. X. (1995). *Economic Growth*, McGraw Hill, New York.

Bleaney, M. and Nishiyama, A. (2002). 'Explaining growth: a contest between models', *Journal of Economic Growth*, Vol. 7, pp. 43–56.

Breusch, T. S. and Pagan, A. R. (1980). 'The Lagrange multiplier test and its application to model specification in econometrics', *Review of Economics Studies*, Vol. 47, pp. 239–253.

Breusch, T. S. (1990). 'Simplified extreme bounds', in Granger C. W. J. (ed.), *Modelling Economic Time Series: Readings in Econometric Methodology*, Clarendon Press, Oxford, pp. 72–81.

Brock, W. A. and Durlauf, S. N. (2001). 'Growth empirics and reality', *World Bank Economic Review*, Vol. 15, pp. 229–272.

Brownstone, D. and Kazimi, C. (1998). *Applying the Bootstrap*, Unpublished Typescript, University of California, Irvine.

Chow, G. C. (1960). 'Tests of equality between sets of coefficients in two linear regressions', *Econometrica*, Vol. 28, pp. 591–605.

DeLong, J. B. and Summers, L. H. (1991). 'Equipment investment and economic growth', *Quarterly Journal of Economics*, Vol. 106, pp. 445–502.

Doppelhofer, G., Miller, R. I. and Sala-i-Martin, X. X. (2000). *Determinants of Long-term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach*, National Bureau of Economic Research Working Paper No. 7750, June.

Ericsson, N. R., Campos, J. and Tran, H.-A. (1990). 'PC-GIVE and David Hendry's econometric methodology', *Revista de Econometrica*, Vol. 10, pp. 7–117.

Faust, J. and Whiteman, C. H. (1995). Commentary [on Grayham E. Mizon ''Progressive modeling of macroeconomic times series: The LSE methodology'']', in Hoover K. D. (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, Kluwer, Boston, pp. 171–180.

Faust, J. and Whiteman, C. H. (1997). 'General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: a translation and critique', *Carnegie-Rochester Conference Series on Economic Policy*, Vol. 47, pp. 121–161.

Fernandez, C., Ley, E. and Steel, M. F. J. (2001). 'Model uncertainty in cross-country growth regressions', *Journal of Applied Econometrics*, Vol. 16, pp. 563–576.

Gilbert, C. L. (1986). 'Professor Hendry's econometric methodology', *Oxford Bulletin of Economics and Statistics*, Vol. 48, pp. 283–307.

Granger, C. W. J. and Uhlig, H. F. (1990). 'Reasonable extreme-bounds analysis', *Journal of Econometrics*, Vol. 44, pp. 159–170.

Grier, K. B. and Tullock, G. (1989). 'An empirical analysis of cross-national economic growth', *Journal of Monetary Economics*, Vol. 24, pp. 259–276.

Hansen, B. E. (1996). 'Methodology: alchemy or science?', *Economic Journal*, Vol. 106, pp. 1398–1431.

Hendry, D. F. (1987). 'Econometric methodology: a personal viewpoint', in Bewley T. (ed.), *Advances in Econometrics*, Vol. 2. Cambridge University Press, Cambridge.

Hendry, D. F. (1988). 'Encompassing', *National Institute Economic Review*, No. 125, August, pp. 88–92.

Hendry, D. F. (1995). *Dynamic Econometrics*, Oxford University Press, Oxford.

Hendry, D. F. (1997). 'On congruent econometric relations: A comment', *Carnegie-Rochester Conference Series on Public Policy*, Vol. 47, 163–190.

Hendry, D. F. and Krolzig, H.-M. (1999). 'Improving on ''data mining reconsidered'' by K. D. Hoover and S. J. Perez', *Econometrics Journal*, Vol. 2, pp. 202–218.

Hendry, D. F. and Krolzig, H.-M. (2001). *Automatic Econometric Model Selection using PcGets 1.0*, Timberlake Consultants, London.

Hendry, D. F. and Krolzig, H.-M. (2003). *Sub-sample Model Selection Procedures in Gets Modelling*, Nuffield College, Oxford Discussion Paper No. 2003-W17, April.

Hendry, D. F. and Mizon, G. E. (1990). 'Procrustean econometrics: Or stretching and squeezing data', in Granger C. W. J. (ed.), *Modelling Economic Time Series: Readings in Econometric Methodology*, Clarendon Press, Oxford, pp. 121–136.

Hendry, D. F. and Richard, J.-F. (1987). 'Recent developments in the theory of encompassing', in Cornet B. and Tulkens H. (eds), *Contributions to Operations Research and Economics: The Twentieth Anniversary of Core*, Cambridge, MA, MIT Press.

Hendry, D. F., Leamer, E. E. and Poirier, D. J. (1990). 'The *ET* dialogue: A conversation on econometric methodology', *Econometric Theory*, Vol. 6, pp. 171–261.

Hoover, K. D. (1995). 'In defense of data mining: some preliminary thoughts', in Hoover K. D. and Sheffrin S. M. (eds), *Monetarism and the Methodology of Economics: Essays in Honour of Thomas Mayer*, Edward Elgar, Aldershot, pp. 242–258.

Hoover, K. D. and Perez, S. J. (1999). 'Data mining reconsidered: Encompassing and the general-to-specific approach to specification search', *Econometrics Journal*, Vol. 2, pp. 1–25.

Hoover, K. D. and Perez, S. J. (2000). 'Three attitudes towards data-mining', *Journal of Economic Methodology*, Vol. 7, pp. 195–210.

Horowitz, J. L. (1997). 'Bootstrap methods in econometrics: Theory and numerical performance', in Kreps R. and Wallis K. (eds), *Advances in Economics and Econometrics: Seventh World Congress of the Econometric Society*, Vol. 3, pp. 188–222, Cambridge University Press, Cambridge.

Jones, C. I. (1998). *Introduction to Economic Growth*, Norton, New York.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). 'Analyzing incomplete political science data: An alternative algorithm for multiple imputation', *American Political Science Review*, Vol. 95, pp. 49–69.

Kormendi, R. C. and Meguire, P. G. (1985). 'Macroeconomic determinants of growth', *Journal of Monetary Economics*, Vol. 16, pp. 141–163.

Krolzig, H.-M. and Hendry, D. F. (2001). 'Computer automation of general-to-specific model selection procedures', *Journal of Economic Dynamics and Control*, Vol. 25, pp. 831–866.

Leamer, E. E. (1983). 'Let's take the con out of econometrics', *American Economic Review*, Vol. 73, pp. 31–43.

Leamer, E. E. (1985). 'Sensitivity analysis would help', *American Economic Review*, Vol. 75, pp. 308–313.

Leamer, E. E. and Leonard, H. (1983). 'Reporting the fragility of regression estimates', *Review of Economics and Statistics*, Vol. 65, pp. 306–317.

Levine, R. and Renelt, D. (1992). 'A sensitivity analysis of cross-country growth regressions', *American Economic Review*, Vol. 82, pp. 942–963.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Lucas, R. E., Jr (1988). 'On the mechanics of economic development', *Journal of Monetary Economics*, Vol. 22, pp. 3–42.

Lovell, M. C. (1983). 'Data mining', *Review of Economics and Statistics*, Vol. 65, pp. 1–12.

Mankiw, N. G., Romer, D. and Weil, D. N. (1992). 'A contribution to the empirics of economic growth', *Quarterly Journal of Economics*, Vol. 107, pp. 407–437.

McAleer, M. (1994). 'Sherlock Holmes and the search for truth: a diagnostic tale', *Journal Economic Surveys*, Vol. 8, pp. 317–370.

McAleer, M., Pagan, A. and Volcker, P. A. (1985). 'What will take the con out of econometrics', *American Economic Review*, Vol. 75, pp. 293–307.

Mizon, G. E. (1977). 'Model selection procedures', in Artis M. J. and Nobay R. A. (eds), *Studies in Modern Economic Analysis*, pp. 97–120, Blackwell, Oxford.

Mizon, G. E. (1984). 'The encompassing approach in econometrics', in Hendry D. F. and Wallis K. F. (eds), *Econometrics and Quantitative Economics*, Blackwell, Oxford, pp. 135–172.

Mizon, G. E. (1995). 'Progressive modelling of macroeconomic time series: the LSE methodology', in Hoover K. D. (ed.), *Macroeconometrics: Developments, Tensions and Prospects*, Kluwer, Boston, pp. 107–170.

Mizon, G. E. and Richard, J.-F. (1986). 'The encompassing principle and its application to testing non-nested hypotheses', *Econometrica*, Vol. 54, pp. 657–678.

Pagan, A. (1987). 'Three econometric methodologies: a critical appraisal', *Journal of Economic Surveys*, Vol. 1, pp. 3–24.

Phillips, P. C. B. (1988). 'Reflections on econometric methodology', *Economic Record*, Vol. 64, pp. 334–359.

Raftery, A., Madigan, D. and Hoetig, J. A. (1997). 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association*, Vol. 92, pp. 179–181.

Romer, P. M. (1986). 'Increasing returns and long-run growth', *Journal of Political Economy*, Vol. 94, pp. 1002–1037.
Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
Rubin, D. B. (1996). 'Multiple imputation after 18+ years', *Journal of the American Statistical Association*, Vol. 91, pp. 470–489.
Sachs, J. D. and Warner, A. M. (1995). 'Economic reform and the process of economic integration', *Brookings Papers on Economic Activity*, Vol. 1, pp. 1–95.
Sachs, J. D. and Warner, A. M. (1996). *Natural Resource Abundance and Economic Growth*, Unpublished Typescript, Harvard Institute International Development.
Sala-i-Martin, X. X. (1997a). 'I have just run two million regressions', *American Economic Review*, Vol. 87, pp. 178–183.
Sala-i-Martin, X. X. (1997b). *I have Just Run Four Million Regressions*, Unpublished Typescript, Economic Department, Columbia University.
Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, London and New York, Chapman and Hall.
Solow, R. M. (1956). A contribution to the theory of economic growth, *Quarterly Journal of Economics*, Vol. 70, pp. 65–94.
Temple, J. (1998). 'Equipment investment and the Solow model', *Oxford Economic Papers*, Vol. 50, pp. 39–62.
Temple, J. (2000). 'Growth regressions and what the textbooks don't tell you', *Bulletin of Economic Research*, Vol. 52, pp. 181–205.
Wu, C. F. J. (1986). 'Jackknife, bootstrap and other resampling methods in regression analysis', *Annals of Statistics*, Vol. 14, pp. 1261–1295.

Appendix A: The general-to-specific search algorithm used in the simulations

A. The data are generated according to the simulated equation setup with either 0, 3, 7, or 14 true variables included. *Candidate* variables include a constant and all variables in Levine and Renelt's dataset, with the exceptions noted in the main text and in the notes to Memorandum 1, downloadable from our websites: http://www.econ.ucdavis.edu/faculty/kdhoover/research.html and http://www.csus.edu/indiv/p/perezs/Data/data.htm/. A *replication* consists of creation of a simulated dependent variable using one of the simulated models and one draw from the bootstrapped random errors. *Nominal size* governs the conventional critical values used in all of the tests employed in the search: it is 5%. Two overlapping sub-samples are created, each comprising 90% of the data set. Independent searches are run on the two subsamples. A *general specification* is estimated on a replication using a full set of candidate variables.

B. Five search paths are examined. Each path begins with the removal of one of the candidate variables with the five lowest *t*-statistics in the current general specification. All *t*-statistics are calculated using White's heteroscedasticity-corrected standard errors. The first search

begins by re-estimating the regression. This re-estimated regression becomes the *current specification*. The search continues until it reaches a *terminal specification*.

C. The current specification is estimated and all searchable variables are ranked according to their *t*-statistic. The searchable variable with the lowest *t*-statistic is removed.

D. Each current specification is subjected to the following battery of tests:
   i. subsample stability test: an *F*-test for the equality of the variances of the first half versus the second half of the sample. (This is analogous to a Chow test in a time-series context.) This test compares the regressions over each subsample to the regression over the full sample. If the degrees of freedom do not permit splitting the sample into equal subsamples, the test is replaced by one that compares a regression over the first $k + (n - k)/2$ observations to the one over the full sample (on both tests, see Chow, 1960).
   ii. An *F*-test of the hypothesis that the current specification is a valid restriction of the current general specification.

E. The number of tests failed is recorded and the new specification becomes the current specification. Return to C until all remaining variables have a significant *t*-statistic.

F. If all variables are significant, and all of the tests in the test battery are passed, the current specification is the terminal specification and go to H. If any of the tests fails return to the last specification for which all the tests are passed and go to G.

G. The variable with the lowest *t*-statistic is eliminated. The resulting current specification is then subjected to the battery of tests.
   i. If the current specification fails any one of these tests, the last variable eliminated is replaced, and the current specification is re-estimated eliminating the variable with the next lowest insignificant *t*-statistic.
   ii. If the current specification passes all tests, re-estimate and return to G.
   iii. The process of variable elimination ends when a current specification passes the battery of tests and either has all variables significant or cannot eliminate any remaining insignificant variable without failing one of the tests.

H. After a terminal specification has been reached, it is recorded and the next search path is tried until all have been searched.

I. Once all search paths have ended in a terminal specification, the *final specification* is chosen through a sequence of encompassing tests. We form the non-redundant joint model from each of the different terminal

specifications; take all candidate specifications and perform the *F*-test for encompassing the other specifications. If only one specification passes, it is the final specification. If more than one specification passes, the specification with the minimum Schwarz criterion is the final specification. If no model passes, reopen the search on the non-redundant joint model (including testing against the general specification) using only a single search path and take the resulting model as the final specification.

J. The final specification is the intersection of the two specifications from each subsample.

Appendix B: Multiple imputation procedures

All the regressions in both the general-to-specific and modified extreme bounds approaches in section IV use multiple imputation to fill in the missing values in the data set. Multiple imputation is unfamiliar to many economists. Standard references include Little and Rubin (1987), Rubin (1987, 1996), and Schaefer (1997). We have followed closely the procedures advocated by King *et al.* (2001).

Multiple imputation has two steps. First, missing values are treated as parameters and the maximum likelihood estimate of the distribution of each is formed on the basis of all of the available data. The imputed value is drawn from the estimated distribution using a procedure similar to bootstrapping. Secondly, in order not to introduce spurious precision into estimates based on imputed values, repeated draws are used to construct multiple data sets. Any regressions are run on each data set and the coefficient estimates and standard errors combined to form a joint estimate. Monte Carlo studies of multiple imputation suggest that it maintains size, but loses some power relative to complete data sets. King *et al.* show that it is superior to casewise deletion in almost all realistic applications.

The essential step is to regard missing data as parameters to be estimated through maximizing a likelihood function. Theoretically, the imputation–posterior (IP) method converges to an exact distribution. Unfortunately, the method converges slowly in distribution only, and there is no mechanical convergence criterion, so that considerable judgment is required. King *et al.* propose a fast, robust alternative: the expectations–maximization algorithm with importance sampling (EMis). They demonstrate in a Monte Carlo study that it does well in closely approximating the IP algorithm and in recovering the true coefficients and standard errors when there is missing data. The algorithm is implemented in a program called *Amelia* downloadable from http://gking.harvard.edu.

Following King *et al.* (2001) we start with the entire data set (62 variables) with missing values and draw five different values for each missing value to create five complete sets. Each regression in both the general-to-specific and the extreme-bounds analysis is then conducted separately on each of the five data sets and the results combined according to the formulae given in King *et al.* (2001, p. 53). The EMis algorithm requires the number of observations in the data set to exceed $p(p + 3)/2$, where p is the number of variables. With 62 variables, at least 2,015 observations would be needed for the algorithm to work. We have only 138. In such cases, the manual for *Amelia* recommends the addition of ridge priors (see Amemiya, 1985, chapter 2, section 2). In our case, the effect would be to add a large number of lines of random observations to the data matrix. The effect at the point of imputation would be to replace the conditional distribution for the variable to be imputed with an unconditional one. We, therefore, do this directly by drawing each imputed value from a normal distribution with a mean equal to the sample mean of the available observations on that variable and a variance equal to the sample variance. Informal simulations comparing results from complete data sets (the same trimmed Levine and Renelt data set used in the simulations in section III) to the same sets with missing data gave good results.

Again following King *et al.*'s advice bounded and/or asymmetrical variables are transformed to approximate unbounded or, at least more symmetrical, variables before imputation. After imputation the transformations are inverted. Integer-valued variables are either rounded to the nearest integer or, in some cases, the inverted value is used to set the probabilities of a uniform bivariate random distribution from which the imputed value is drawn. The transformations for each variable are described in Memorandum 2, downloadable from our websites: http://www.econ.ucdavis.edu/faculty/kdhoover/research.html and http://www.csus.edu/indiv/p/perezs/Data/data.htm.