Bulletin of Economic Research 52:3, 2000, 0307-3378

GROWTH REGRESSIONS AND WHAT THE TEXTBOOKS DON'T TELL YOU*

Jonathan Temple
Nuffield College and Institute of Economics and Statistics,
University of Oxford

ABSTRACT

The paper discusses three econometric problems that are rarely given adequate discussion in textbooks: model uncertainty, parameter heterogeneity, and outliers. Leamer's extreme bounds analysis can be adapted to address all three problems simultaneously. Two examples are presented based on an influential cross-country growth paper by Levine and Renelt (*American Economic Review*, vol. 82, 1992, pp. 942–63).

I. INTRODUCTION

My purpose in writing this paper is to draw attention to three econometric problems which are rarely discussed, but are often crucial in certain kinds of applied work. The three problems are model uncertainty, parameter heterogeneity, and outliers. I will illustrate the surrounding difficulties using examples drawn from the cross-country growth literature, partly because this is the area of applied econometrics I am most familiar with, and partly because all three problems are likely to be particularly important in this context.

It is important to emphasize that these problems are all well-known to statisticians, and much of what I have to say is not new. Yet it is perhaps fair to say that most applied economists are rather less familiar with the

* I am grateful to Jouni Kuha, John Muellbauer, Steve Nickell, Steve Redding and Mark Steel for useful discussion of the issues raised here, and to Ross Levine for supplying data. An anonymous referee and an editor provided very helpful comments. Any errors are my responsibility.

issues than statisticians, not least because the relevant material is largely missing from mainstream econometrics textbooks. This article attempts to fill the gap. In doing so, it deliberately falls somewhere between a conventional research report and something more like a textbook chapter.

The main original contribution is to modify an existing approach to model uncertainty, Leamer's extreme bounds analysis, so that it addresses some possible criticisms of Leamer's ideas. I then illustrate the use of the modified approach, using two examples based on the influential paper of Levine and Renelt (1992). The results on the determinants of growth should be of independent interest.

In the remainder of this introduction, I provide a brief introduction to each of the three problems, and then describe how the paper seeks to address them. The first problem to be considered is in some ways the most fundamental, and controversial: model uncertainty. The starting point is a simple proposition: the aim of most applied research in economics is to communicate the degree of support for one or more hypotheses about unknown parameters. It is generally accepted that, as well as presenting estimates of the parameters, researchers should also provide information on the likely reliability of the estimates. Hence it is common to report a preferred model accompanied by the results of diagnostic tests, usually indicating that the model is unlikely to be flawed in a variety of respects.

Over the last twenty years or so, this approach has started to attract searching criticism. The central difficulty is that several different models may all seem reasonable given the data, but lead to very different conclusions about the parameters of interest. In these circumstances, presenting the results of a single preferred model can often be misleading. It leads the reader to underestimate the uncertainty actually present about the parameters, because the reader is rarely made aware of the inevitable uncertainty surrounding the form of the model.

This problem has not gone unnoticed. Leamer (1978, 1983, 1985) and Leamer and Leonard (1983) have suggested a technique they call 'extreme bounds analysis' (EBA), which is essentially a means of reporting the sensitivity of results to changes in specification. It has only rarely been applied in the literature. In Section II, I discuss why this might be the case. I argue that the most common objections to EBA are either misplaced or easily taken into account by a careful presentation of the results.

Another aim of the paper is to show how EBA can be extended to address two other important problems. The first of these is parameter heterogeneity. Say that, as is usual in the growth literature, we estimate a regression using a set of cross-sectional units, perhaps regions or countries. Then an implicit assumption in the regression analysis is that the parameters are constant across units, or in other words that the countries all follow the same underlying model relating growth to its determinants.

If we are interested only in estimating the parameter averages, we can weaken this slightly, so that we assume only that parameters are distributed independently of the variables in the regression. Yet even this weaker assumption is often likely to be too strong. In the growth literature it is easy to suggest examples of parameters that are likely to be correlated with variables in the regression. For instance, political instability may be associated with both lower investment, and a lower impact of investment on growth, so that the coefficient on investment should ideally be allowed to vary across countries. It would not be difficult to construct further examples.

A sensible response to this is to find ways of modelling the heterogeneity. For instance, in the example above, it might be possible to reduce the extent of heterogeneity by introducing an interaction term between political instability and investment. At some point, though, even the most careful researcher has to accept that any useful model embodies certain restrictions on parameters. At least in economics, this means that almost all our empirical models should be regarded as approximations, albeit potentially useful ones.

Some statisticians argue that, once this point is acknowledged, it is useful to consider alternatives to estimation by least squares. This paper will make the case for using robust estimation techniques, including methods which are now available in standard computer software. In some circumstances, these techniques may help address the heterogeneity problem. Certain robust estimators can be thought of as trying to seek out the most coherent part of the data, the part best approximated by the model being estimated. This interpretation should not be pushed too far, but it does indicate a closely related advantage of robust estimation. These estimators will not often be led astray by outliers, the third and final econometric problem to be considered.

The usual presentation of outliers emphasizes that one or more observations may be measured with a substantial degree of error, perhaps because of a coding or transcription mistake. When we take a parameter heterogeneity perspective, it is clear that we could think about outliers in another way. Some observations may be entirely correct, but drawn from a different regime. Our constant parameter approximation implicitly assumes only one regime generates the data, or that the parameters in the different regimes vary randomly. As a result, observations for which the model is a particularly bad approximation may end up exerting an unduly large influence on the results.

Again, the use of robust estimation is a sensible response. One aim of this paper will be to demonstrate how robust estimation may be combined with an extreme bounds analysis. This allows the applied researcher to address simultaneously the problems of model uncertainty and outliers, and to a lesser extent, perhaps parameter heterogeneity as well. The method is applied to the dataset used in the influential paper by Levine and Renelt (1992). This paper has been widely cited in the empirical growth literature, and I show how their results change when robust methods are applied.

Before then, Section II discusses a variety of approaches to model uncertainty, placing special emphasis on extreme bounds analysis. Section III introduces parameter heterogeneity and outliers. Section IV examines proposals for improving extreme bounds analysis, and sets out a suggestion based on robust estimation. This method is then illustrated in Section V, using two examples drawn from the cross-county growth literature. Section VI concludes.

II. APPROACHES TO MODEL UNCERTAINTY

The problem of model uncertainty has gained increasing attention in the statistics literature. Useful and readable introductions can be found in Chatfield (1995) and Raftery (1995), who both discuss the problems that model uncertainty poses for classical inference. In this section, I discuss two approaches to model uncertainty that have been proposed: extreme bounds analysis and Bayesian model averaging.

II.1. Extreme bounds analysis

The idea that model uncertainty should be taken into account when presenting results is an easy one to grasp. Despite this, the responses proposed in the literature have sometimes met with criticism, at least from econometricians. In what follows, I set out the method of extreme bounds analysis (EBA), and discuss the usual criticisms of this approach.

The central idea of EBA is to report an upper and lower bound for parameter estimates, thereby indicating sensitivity to the choice of specification. Say that one has a large set of explanatory variables, X. The upper and lower bounds are based on a set of regressions using different subsets of the variables in X, and sometimes the general model which contains all the variables. This seems a relatively natural and transparent approach to the problem of assessing the consequences of model uncertainty, but there appears to be a widespread lack of enthusiasm for the technique.[1]

One of the strongest objections to EBA is that the models generating the bounds might be flawed. Hence the EBA will tend to overstate the degree of uncertainty about parameters, because it ignores the information we have that certain models are poor and should almost

[1] It is criticized by, for instance, McAleer *et al.* (1985) and Hendry and Mizon (1990).

certainly be dismissed. For instance, say we have prior knowledge that one of the variables in X should be present in the model. When we examine sensitivity to the choice of subsets of X, this variable will sometimes be omitted, and the models omitting this key variable may be precisely the ones that generate wide intervals for the parameters.

What this suggests is that we should think carefully about how the results of an EBA should be presented. Rather than simply reporting the upper and lower bounds, it is often possible to present a table listing a variety of models, together with the results of diagnostic tests for each model. The reader can then assess the strength of support for various hypotheses, taking into account their own views about what would constitute an acceptable model.

As McAleer (1994) points out, most applications of EBA do not provide enough evidence of the usefulness of their estimated models. Criticizing the study of Levine and Renelt (1992), he writes that 'any inferences regarding fragility which ignore functional form misspecification and other departures from the standard conditions should be seriously called into question ... since the only summary statistic reported [in Levine and Renelt] is the coefficient of determination, and no diagnostic tests whatsoever are reported, it is hard not to withhold judgment regarding the statistical "significance" of the extreme bounds or of *any* OLS estimates of the regression parameters' (pp. 349–50).

When the number of variables in X becomes large, as it is in Levine and Renelt (1992), a complete presentation of the results becomes unwieldy, and we must explore alternatives. The simplest is to present diagnostic tests for the particular regressions which generate the upper and lower bounds. If these are satisfactory, the bounds should be regarded with less scepticism.

The problem remains that an important variable may be omitted. Here it is worth noting that most applications of EBA partition variables into a set that are present in each regression, and a set of 'doubtful' ones that are tried in turn and in different combinations. This partition is important for the results, but it is sometimes argued that the initial partition is likely to be rather arbitrary. However, there is no reason why standard model selection procedures (such as testing down from a general specification) cannot be used in advance to identify variables which seem to be particularly important.

At a more theoretical level, it can be shown that the results of an EBA do not contain any information beyond that in χ^2 statistics for variable deletion. Thus, McAleer *et al.* (1985) argue that selectivity in reporting such statistics has, as an exact analogue, the arbitrary classification of variables when carrying out the EBA. They propose that it would be simpler to present estimates of a general model, together with χ^2 statistics that allow readers to judge for themselves whether further simplification is justified.

There are two difficulties with this proposal. The first and most obvious is that nonspecialist readers will rarely have the experience and expertise to make these judgements sensibly. In a sense, the proposal misses the point of EBA: why substitute a list of rather opaque χ^2 statistics for at least some explicit information about alternative models? As Breusch (1985) points out, one view of the EBA is that it presents information on sensitivity to the choice of model in a form which is particularly easy to understand and assimilate. Along similar lines, Pagan (1987) concluded that whenever several variables are deleted in a simplification exercise, provision of the extreme bounds for the parameters of interest is desirable.

A second difficulty is the underlying assumption that the general model is the most useful starting point. It is quite common in economic applications, and particularly in studying growth, to have a large set of possible explanatory variables, at least relative to the number of observations. One consequence is that presenting the general model will often be rather uninformative.

The conventional response, implicit in most criticisms of EBA, is that researchers should use well-known techniques to 'test down' to a preferred specification. I have already discussed one problem inherent in such an approach. Presenting a single preferred model can lead the reader to underestimate the degree of uncertainty surrounding the parameter estimates.

This might be less of a problem if we could establish a model which clearly dominated all others across all dimensions. We might then be justified in regarding this model as the best available approximation to the truth. Unfortunately, it is rarely possible to establish that a model does dominate other possibilities in this way. Model selection is an art rather than a science, and even very careful and systematic approaches to arriving at a final model usually leave open the possibility that discarded alternatives are in fact superior.

Consider the choice of a more parsimonious model, for instance by successively eliminating the least significant variables. One problem here is that p-values or t-statistics based on a model selected from a large set of possibilities should not be interpreted in the same way as when only two models were ever considered. Granger *et al.* (1995) point out other potentially fundamental problems with the use of hypothesis testing in model selection. Testing favours the null hypothesis, and typically uses a significance level that is arbitrary. Starting from the same data, researchers can arrive at different final models, depending on the order in which tests are carried out and the significance levels used.

In summary, it is rare that we can say with certainty that our preferred model dominates all other possibilities in all dimensions. In these circumstances, it makes sense to provide some information about how sensitive the findings are to alternative modelling choices. Extreme bounds analysis provides a relatively simple means of doing exactly this.

A final objection to EBA is that it is too mechanical, and diverts attention from more important tasks. It is certainly true that an EBA should not be regarded as a substitute for the procedures recommended by econometricians. For instance, it remains very important to give thorough consideration to such things as diagnostic testing and the specification of functional form, and hence identify classes of acceptable models. Given how quickly an EBA can be carried out, however, it is hard to see that its use will lead to an overall decline in the quality of modelling efforts.

II.2. Other approaches to model uncertainty

It should be noted that there are other responses to model uncertainty, besides extreme bounds analysis. One approach gaining increasing attention in the statistics literature is 'Bayesian model averaging' (BMA). The idea is to average over a variety of alternative models. Priors are attached to the models as well as to the parameters; based on the data, posterior probabilities for the different models are calculated. Distributions for the parameters can be calculated by averaging the posterior distributions under each of the models considered, weighted by their posterior model probability. Chatfield (1995), Draper (1995) Raftery *et al.* (1997) and Hoeting *et al.* (1999) provide readable introductions to these methods.

It is interesting to note that an important inspiration for this approach is Leamer (1978), who provided the basic framework for the Bayesian approach to model uncertainty. As Hoeting *et al.* (1999) note, difficulties in implementation meant that the averaging approach has been neglected until recently. Hence where model uncertainty has been addressed, it is typically by using what might be regarded as watered-down versions of these ideas, either Leamer's extreme bounds analysis or a related sensitivity analysis.

Theoretical advances and improvements in computational power have allowed researchers to develop approaches that conform more closely to Leamer's original framework. It is clear that this is potentially a very important development. One reason for its strong appeal is that the weights used in the final averaging procedure are tied quite closely to the predictive ability of the different models. Hence the models are weighted in a way that is less arbitrary than more traditional and informal approaches to model uncertainty. Instead, the averaging process is based on a sensible index of model adequacy. (See Geweke (1999) for a more precise discussion.)

Ley and Steel (1999) have recently applied the methods in the context of growth regressions, using the same dataset as Sala-i-Martin (1997), with some very intriguing findings. One conclusion is that using BMA to average over growth models usually leads to better out-of-sample predictions than a 'null' model with entirely random variation. Although this may seem unsurprising, it is rather important, given the occasional claim that growth regressions are likely to be entirely spurious.

The remaining problems inherent in implementing BMA are reviewed by Hoeting *et al.* (1999). Some are technical, to do with convergence issues in the use of Markov-chain Monte Carlo methods to average across models with many predictors. Open questions include the choice of prior distribution for the different models, the choice of the class of models to average over, and the performance of the procedure when the true model is not in the chosen class.

These open questions are arguably crucial in the growth context, and suggest that BMA is unlikely to provide a complete approach by itself. For instance, Ley and Steel follow earlier writers in averaging over a restricted set of models, those that are linear in the predictors. Someone following a more conventional *ad hoc* approach might discover through diagnostic testing the potential importance of, say, an interaction term. This model may well turn out to make better out-of-sample predictions than the BMA procedure. It is difficult or perhaps impossible to give the orthodox approach any formal grounding; but an eclectic mix of Bayesian averaging and conventional approaches to model-building might well be justified on pragmatic grounds.

The precise details of applying BMA remain somewhat controversial. One key point to emphasize is that mixing over a variety of models is a separate step from generating their posterior probabilities. The latter element of BMA may have a wider appeal than any subsequent mixing stage. In this respect it is interesting to note the positions of Chatfield (1995) and Sir David Cox: Chatfield, though generally supportive, notes that one disadvantage of BMA is that it does not lead to a simple model, which can make interpretation of the results harder. Sir David Cox, in the discussion of Draper (1995), argues that it is sometimes more sensible for researchers to report a variety of models than average over a wide class. One justification for this might be that, if a variety of models are reported, readers can bring to bear their own priors on what would constitute a sensible model.

For all these reasons, there is still a case for using extreme bounds analysis in some contexts. One advantage retained by the extreme bounds approach is that one can investigate the models generating the bounds. For instance, if these particular models perform poorly on diagnostic tests, one may want to reject them, allowing a narrowing of the bounds. Presentation of different models allows the reader to draw his or her own conclusions about the degree of uncertainty surrounding the choice of model and the parameter estimates, and this is the approach followed below.

None of this is to deny that Bayesian model averaging is likely to provide an increasingly important tool. It reflects Leamer's original ideas much more closely than extreme bounds analysis. More precisely, when all plausible models are contained in the averaged class, it is the formally correct approach to model uncertainty from a Bayesian perspective. That said, I believe that one should not discard more conventional methods altogether, for the reasons set out above. The best approach to issues like diagnostic testing is an open question, and further discussion can be found elsewhere, particularly in Box (1980) and Geweke (1999).

III. PARAMETER HETEROGENEITY AND OUTLIERS

I hope to have indicated that extreme bounds analysis can provide a useful complement to other methods of investigation, especially if the presentation is done carefully. Before moving on to an illustration based on growth regressions, this section discusses some of the issues surrounding parameter heterogeneity and outliers. These issues are likely to be particularly important in cross-country growth regressions, but it should be noted that they arise in other areas. For instance, labour economists often estimate returns to schooling, and parameter heterogeneity will also be very important in that context.

III.1. Parameter heterogeneity

At one level, it may seem strange to give special emphasis to 'parameter heterogeneity', since it is just one aspect of regression misspecification. My reason for drawing attention to the problem is that it appears to be much less frequently considered by applied researchers than better-known difficulties such as endogeneity or measurement error.

This lack of consideration can be seen in the way growth researchers choose to deal with heteroskedasticity. It is common to use heteroskedasticity-consistent standard errors, and not pay further attention to the problem. Yet heteroskedasticity may indicate parameter heterogeneity, which could have serious implications for the consistency of parameter estimates. To see this, note that Chesher (1984) derives a test for random parameter variation which, for linear models, turns out to be equivalent to a combination of widely used tests for heteroskedasticity and non-normality (see Hall, 1987).

The possible connection here can be indicated in other ways. One of the tests for heteroskedasticity derived by White (1980) involves a regression of the squared residuals on all squares and cross-products of the regressors. To some extent, the inclusion of cross-products can be seen as testing for omitted interaction terms, and hence this is best interpreted as a joint test for heteroskedasticity and misspecification.[2] If the test gives unsatisfactory results, this may indicate that the model should be respecified in such a way that parameter heterogeneity is lessened. For similar reasons, other specification tests, such as Ramsey's RESET approach, could also be used far more routinely than they are at present. A test specifically designed to detect parameter heterogeneity is provided by Dutta and Leon (1991).

Parameter heterogeneity has not been ignored altogether, and Durlauf and Johnson (1995) in particular have highlighted it in the growth context. Other researchers have considered the effects in the context of dynamic panel data analyses and growth regressions (Pesaran and Smith, 1995; Lee *et al.*, 1997). For the most part, though, applied researchers rarely take the problem into account.

There is a practical consequence for the use of extreme bounds analysis. When presenting the regression models that generate the extreme bounds for the coefficients of interest in an EBA, it is also useful to report the results of diagnostic tests that may pick up misspecification. To put it at its simplest, any extreme bound based on a model that appears to be misspecified should not be trusted, at least from a classical perspective. This can help narrow the bounds.

III.2. Outliers

Another reason for emphasizing parameter heterogeneity is that it draws attention to the potential importance of outliers. In cross-section or panel data econometrics, it is particularly important to ensure that results reflect what is going on in the majority of the sample. It is entirely possible that some observations are drawn from a different regime, that is one with different parameters. If we want to draw useful generalizations, we do not want the results to be driven by a minority of observations drawn from a different regime.

It helps to start with some terminology. A 'response outlier' occurs when the dependent variable takes on a value that is unusual, given the explanatory variables. A 'design outlier' is a set of values for the explanatory variables which is unusual, and at some distance from the remainder of the data. In the context of OLS, it is often referred to as a *leverage point*, since it can exert a large influence on the OLS estimates. Sometimes, the response variable takes a value near that predicted by the bulk of the data, and the presence of a leverage point will tend to raise the precision of the coefficient estimates.[3] A leverage point is bad news

[2] The interpretation in terms of omitted interaction terms is not exact, since the heteroskedasticity test is based on the *squared* residuals.

[3] Ruppert and Simpson (1990) argue that this increased precision is sometimes misleading. Given the likelihood of measurement errors, one might want to restrict the influence of leverage points.

when the corresponding response is outlying. In this case, the leverage point will tend to pull the OLS estimates towards it, distorting the results.

In the unlikely event that our model exactly reflects the structure of the data generating process, the outlier problem can be regarded as simply an efficiency issue. Even then, it is worth remembering that, when the underlying disturbances are not normally distributed, OLS is no longer best unbiased. It is still the best linear estimator, but without normal errors, *any* linear estimator will often perform extremely poorly (Hampel *et al.*, 1986). Tests for normality of the residuals are rarely carried out. Even where normality is not rejected, there may still be an underlying problem, since the test has to be carried out on the estimated residuals, not the true disturbances. The OLS estimator will try to make the residuals look as Gaussian as possible, even when the underlying disturbances have a very different distribution.[4]

A deeper way of looking at the problem is to acknowledge that, for most problems faced in economics, our model is only an approximation. It is true that we can sometimes construct a good approximation that accounts not only for the majority of the data, but also for the extreme cases. Dummy variables and transformations can be used, for instance. Even then, it is widely acknowledged that outside certain special cases (perhaps based on physical laws) no model can ever capture reality exactly. In these circumstances it is sometimes useful to complement careful attention to the specification with estimation methods other than least squares, methods that take into account the fact that our model is not exact. Often these methods are particularly useful for identifying the directions in which the model should be extended and improved.

Sometimes the problem of outliers is addressed by dropping each observation in turn, or looking for data points with extreme values of the explanatory variables (design outliers) or omitting observations with high residuals. It is well known in the statistics literature, as discussed below, that all these standard procedures are inadequate. Although clearly preferable to ignoring the problem altogether, such methods are a long way behind current best practice.

Sometimes, ‘single-case’ outlier diagnostics are used: the studentized residuals, Cook’s distance measure, or DFITS statistics. Cook’s distance for the ith observation is a measure of the distance between the coefficient estimates when observation i is included and when it is not. The DFITS statistics are similar but give relatively more weight to leverage points, since they show the effect on an observation’s fitted value when that particular one is dropped from the sample.[5] Other useful

[4] See the commentary on Janson (1988) for this point.

[5] See Atkinson (1985) and Bollen and Jackman (1990) for arguments that DFITS may be superior to Cook’s distance measure.

statistics include DFBETA, which allows one to present a table of changes in the regression coefficients when the *i*th observation is dropped, and similarly DFSTAT, the change in the *t*-statistics. See Atkinson (1985) and Bollen and Jackman (1990) for good introductions, and the text by Belsley *et al.* (1980) for more details.

When these diagnostics are used, or less formal procedures like dropping each observation in turn, it is often as part of a sensitivity analysis after the main results have been obtained. It makes more sense to use them as part of an exploratory analysis, using the information from this to arrive at a better specification. Statistics like Cook's distance should be used in developing a model, given the interaction of outliers with the choice of variables.

Although useful, and often easily available in computer packages, single-case diagnostics like Cook's distance and DFITS are well known to be inadequate in the presence of multiple outliers or leverage points. A group of outliers can mask each other when testing for a single one (the masking effect) or lead to representative observations being wrongly identified as outliers (the swamping effect).[6] Although the single-case measures can be extended, by assessing the effect of dropping several observations at once, this quickly becomes burdensome.

An alternative method is needed, and is provided by robust regression.[7] Two closely related methods seem particularly suitable: least median of squares (LMS) and least trimmed squares (LTS), both due to Rousseeuw (1984). LMS minimizes the median of the squared residuals. LTS typically minimizes the sum of squares over half the observations, the chosen half being the combination which gives the smallest residual sum of squares. The properties of these estimators are not fully understood, but at present, LTS is generally thought preferable to LMS.[8]

One way of thinking about these estimators, particularly least trimmed squares, is that they seek out the portion of the data which is best approximated by a constant parameter model. In the words of Rousseeuw *et al.* (1999, p. 425), the LMS, and LTS methods 'search for a concentrated linear cloud with the majority of the data'. It should be clear that these methods may be particularly well suited to an exercise such as a growth regression, where the idea is to learn about possible

[6] See Davies and Gather (1993), Hadi and Simonoff (1993) and McKean *et al.* (1993) for this point.

[7] See Judge *et al.* (1985), Kennedy (1992) and especially Rousseeuw and Leroy (1987) for an introduction.

[8] For more discussion of these estimators, see Appa and Land (1993), Hettmansperger and Sheather (1992), Rousseeuw (1993), Rousseeuw and Van Zomeren (1990), Ruppert and Simpson (1990), Stromberg (1993) and particularly Rousseeuw *et al.* (1999). Note that the choice of minimizing over half the observations can be altered, and the proportion increased if required.

generalizations in the context of many disparate countries, some of which may be exceptions to the general pattern.

It is important not to push this interpretation too far, because if parameters differ across units, this could give rise to inconsistent estimates in the usual way. It remains true that use of robust estimators may help identify one or more observations that are drawn from a different regime. Differences between the robust and the least-squares estimates may suggest ways in which the model can be respecified, in a way that makes it a better approximation to the majority of the data.

These points are not new. Hampel *et al.* (1986, pp. 4–5) argue that robust estimation can answer several important questions:

> 'Which minorities behave differently, and how? What is the influence of different parts of the data on the final result? Which data are of crucial importance, either for model choice or for final results, and should be examined with special care?'

These questions are of interest in almost any modelling exercise. The fundamental point here was made by one of the founding fathers of robust statistics, in Huber (1981). The use of robust estimators acknowledges that the parametric model will rarely be exact, and takes this into account in a formal way. The consequences of neglecting this point have been forcefully stated by Swartz and Welsch (1986, pp. 169–71):

> 'The Gauss–Markov theorem establishes that for at least one choice criterion (linear, unbiased), OLS is the optimal estimator of the coefficients of the linear model. What is often overlooked in the econometrics literature is that other criteria exist by which to judge estimators. For example, OLS and many other commonly used maximum likelihood techniques have an unbounded influence function; any small subset of the data can have an arbitrarily large influence on their coefficient estimates. In a world of fat-tailed or asymmetric error distributions, data errors, and imperfectly specified models, it is just those data in which we have the least faith that often exert the most influence on the OLS estimates.'

Robust estimators are less efficient than OLS when the disturbances are normally distributed, and in their current form usually lack the same range of diagnostic statistics. Standard errors can be obtained by bootstrapping (Efron and Tibshirani, 1993, pp. 117–21), but a simpler solution is to obtain robust parameter estimates, use the residuals to identify observations that may be problematic, and then drop or downweigh them in an otherwise conventional least-squares regression. This is known as re-weighted least squares (RWLS) and it is recommended by Rousseeuw and Leroy (1987) among others.

One final note is in order. Depending on the number of observations that are dropped in the second stage, the procedure used here might be interpreted as being closer to a sample selection criterion than a check for outlier effects. This alternative interpretation hints at the potential dangers surrounding the method. The idea is to restrict attention to the part of the sample which is most likely to be adequately explained by the model under consideration: but this will tend to mean that even a poor model might fit relatively well in the restricted sample. It is therefore important to note that the exclusion of outliers could be seen as a particularly dangerous form of data mining. To take this into account, Chatfield (1995) recommends carrying out two analyses, both with and without outliers, and that is the approach taken here.

IV. IMPROVING EXTREME BOUNDS ANALYSIS

In this section, I first discuss some improvements to EBA that various researchers have proposed. The remainder of the section discusses a further improvement, which takes into account the interaction between model selection and outliers.

IV.1. Previous improvements

Several authors have introduced improvements to Leamer's original idea. One observation is that the extreme bounds are stochastic, and their point estimates subject to sampling error. The standard errors are difficult to derive analytically, but McAleer and Veall (1989) describe how bootstrapping can be used to obtain them, while Magee (1990) provides a useful asymptotic approximation.

One way of narrowing the bounds is discussed by Granger and Uhlig (1990, 1992). They suggest ignoring any bounds generated by models with an R^2 not within a certain fraction of that achieved by the general model. This should rule out the poorest models, likely to have omitted-variable problems, and this should sometimes narrow the bounds. Sala-i-Martin (1997) proposes an alternative but related method, which is to weight point estimates and variances by the integrated likelihoods of the various models. Hence this technique gives more weight to the outcomes of models which fit relatively well.

Sala-i-Martin notes a difficulty with this general approach, which is that the within-sample fit of a model is sometimes a poor index of its adequacy. For instance, a model may appear to have high explanatory power because one or more of the regressors are endogenous. Once again, this suggests that any approach to model uncertainty should be accompanied by careful attention to the choice of models considered or averaged over.

IV.2. Model uncertainty and outliers

It should be clear by now that outliers, if not carefully dealt with, can play a key role in influencing modelling choices and the final parameter estimates. It is well known in the statistics literature that the presence of a few influential outliers can either hide a relationship, or create the appearance of one where none exists. Hence outliers can sometimes have important consequences for the choice of variables, as discussed by Chatterjee and Hadi (1988).

This suggests that any good approach to model uncertainty should ideally be robust to observations that are measured with error, or drawn from a different regime. In this paper, I propose using a simple variant of EBA in which each regression is first estimated by robust methods. The technique is demonstrated using the dataset of Levine and Renelt (1992), and I show how it can lead to different results.

I carry out the robust EBA by using re-weighted least squares to estimate each regression, having used least trimmed squares to screen for outliers in each regression that is carried out. Since a large number of regressions are required in this kind of analysis, the decision rule for identifying outliers should ideally be simple enough that it is easily programmed into computer software. Here, a country is dropped if its growth rate is two percentage points or more away from that predicted by the LTS regression. This usually means that 10–15 countries are dropped from a typical sample of about 60–100.

Inevitably this is arbitrary to some extent. I have deliberately chosen to omit a relatively high number of observations. This seems safer than omitting only a few, and hence running the risk that results will be influenced by a small cluster of multivariate outliers. The aim is to characterize the most coherent part of the data, and to do this successfully it may be necessary to omit a relatively large number of observations.

Some other potential objections should also be noted. For instance, an informative leverage point may be wrongly identified as a problem. Atkinson (1986) and Fung (1993) suggest using a 'confirmatory' analysis, adding back the supposed outliers one at a time and checking the effect on the estimates. Hadi and Simonoff (1993) argue that the use of LMS or LTS is inferior to other methods when outliers occur at points with low leverage, and is ineffective when there is a high proportion of leverage points in the dataset.

Ruppert and Simpson (1990) argue that using an estimator which always downweights leverage points is prudent, whether or not the response variable is outlying. For reasons like this, more sophisticated approaches may well be preferable to that used here.[9] The rule I have chosen retains one important advantage. It is relatively simple to implement in computer software, a crucial consideration given the number of robust regressions required in an EBA.

[9] More advanced approaches are considered by Rousseeuw and Van Zomeren (1990), Hadi and Simonoff (1993), Atkinson (1994), Ferretti *et al.* (1999), Chang *et al.* (1999), Pena and Yohai (1999) and especially Rousseeuw and Hubert (1999).

V. MODEL UNCERTAINTY AND GROWTH REGRESSIONS

This section discusses the application of extreme bounds analysis in the context of cross-country growth regressions. The results of Levine and Renelt (1992) are discussed, and then compared with some results based on robust estimation.

V.1. Interpretation of Levine and Renelt

The most influential example of an extreme bounds analysis is Levine and Renelt (1992). They carried out an EBA for cross-section growth regressions, and found that many results in the empirical growth literature are not robust to slight changes in the set of conditioning variables. However, growth researchers have also often found outliers to be a major problem. It may be that outlying observations are partly responsible for some of the findings of Levine and Renelt. In what follows, I apply the EBA using robust estimation methods, and contrast some of my results with theirs.

There is a second difficulty with the procedure adopted by Levine and Renelt. As noted by Sala-i-Martin (1997), they use a binary classification of variables, determining whether each is either 'fragile' or 'robust'. A more natural, and in some ways more useful, approach is to try and assign some degree of confidence and interest to each relationship. The results of Ley and Steel (1999) demonstrate the potential usefulness of this kind of approach.

It is worth emphasizing that 'robustness', whether in the Levine–Renelt sense or more generally, is neither a necessary nor sufficient condition for an interesting finding.[10] I will consider both aspects of robustness in turn, taking the 'necessary' aspect first.

If we discover that a variable is not robust to the choice of conditioning variables, this may indicate that we should look deeper, and think carefully about the correlations between the variables and why they might arise. For instance, it might be that a measure of openness does not help explain growth when a measure of investment is included; but a genuine effect of openness may arise precisely through raising investment.

This is a potentially crucial point. A worryingly common interpretation of Levine and Renelt's paper is that only initial income, investment, and human capital matter in the explanation of differing growth experiences, and that nothing else is robust. However, a careful reading makes clear that Levine and Renelt demonstrate not only a robust relationship between growth and the investment ratio, but also between the investment ratio and some indicators of openness to trade (see their Table 9 on

[10] I owe this point to an anonymous referee's comments on Temple (1999).

p. 955). To put this another way, when causality is potentially indirect, a finding that a variable is fragile should be interpreted extremely carefully.

There are other reasons to be wary of 'fragility'. It is worth remembering that the conventional significance levels (1%, 5%, 10%) are determined chiefly by tradition, but embody crucial assumptions about the relative costs of different kinds of testing errors (Leamer, 1978). To see this, note that often the null hypothesis under test is that a variable has no effect. This is tested at a significance level of 10% or less, implying that rejecting the null hypothesis when it is true (a type I error) is relatively costly. In many contexts, it is not clear that this implicit assumption about relative costs is warranted.

To return to the example of growth regressions, researchers will frequently dismiss variables that are sometimes insignificant at the 10% level. Yet this may mean that important relationships come to be neglected by later researchers; in the growth context, it may be type II errors that should be avoided. Tastes differ, but we should certainly not see the 10% level as some kind of fundamental cutoff point in identifying interesting relationships, and should therefore be careful about the interpretation attached to 'fragility'.

What about the importance to be attached to a conclusion in favour of robustness? Is it sufficient for an interesting finding? Attention to the significance of coefficients too often distracts attention from their magnitude. (See McCloskey and Ziliak (1996) for a forceful elaboration of this point.) An effect that is classified as robust and precisely estimated may, on closer inspection, turn out to be of little quantitative importance; or there may be some lurking problem with the econometrics which later calls the result into doubt. Hence we have the overall conclusion that robustness is neither a necessary nor a sufficient condition for a valid and interesting finding.

V.2. Examples of extreme bounds analysis

This paper has advanced several sets of arguments, and it may be useful to recap at this point. Researchers need to find tractable ways of addressing model uncertainty, and extreme bounds analysis remains a potentially useful approach. The results may be more convincing if they take into account the presence of outliers. The presentation of the EBA results should include information about a variety of models, and especially those generating the bounds. Finally, an EBA should be presented in such a way that readers are not led to believe that only robust results are interesting or important.

It may be useful to illustrate these points with some examples.[11] I will focus on two parameters. The first is the coefficient on initial GDP in a

[11] An earlier example may be found in Temple and Johnson (1998), although in that paper less information is provided about the extreme bounds.

growth regression over 1960–89. This is a relatively natural choice, given the widespread attention paid to rates of conditional convergence, which are related to the coefficient on initial GDP. Levine and Renelt classified initial GDP as 'robust'.

The second parameter I consider is the effect on growth of a measure of real exchange rate distortions, over 1974–89. The measure was constructed for ICP benchmark countries and used in some influential growth regressions (Dollar, 1992). Levine and Renelt classify it as 'fragile'.

The variables used to generate the upper and lower bounds are the average ratio of government expenditures to GDP (GOV), the ratio of exports to GDP (X), the average inflation rate (PI), the average growth rate of domestic credit (GDC), the standard deviation of inflation ($STDI$), the standard deviation of domestic credit growth ($STDC$), and an index for the number of revolutions and coups ($REVC$). Between one and three of these variables are included in the regressions. Four variables are always included: initial GDP ($RGDP$), the average investment ratio (INV), the secondary school enrolment ratio (SEC/SED), and population growth ($GPOP$).

The initial GDP results are shown in Table 1. Regressions (1) and (2) report the coefficient and standard error on initial GDP in the two regressions generating the extreme bounds. The table also shows the selection from the extra variables that leads to the bounds, and the p-values of some diagnostic tests.

The first diagnostic test is one for normality of the residuals. The second is a test for heteroskedasticity, called HSW here. It is based on White (1980) and uses an auxiliary regression of the squared residuals on the original regressors and all their squares. A third test, HSF, also stems from White (1980) and involves a regression of the squared residuals on all squares and cross-products of the regressors. (This test cannot be implemented for relatively small samples.) Finally, the Ramsey RESET test of functional form is also reported for each regression.

For each test I report the p-value of the null hypothesis, where the null is that the model is correctly specified. Hence any p-value under 0.05 indicates that the model fails the corresponding diagnostic test at the 5% significance level.

The bare results are not exactly the same as those in Levine and Renelt, but are very close. The message of these results is that the coefficient on initial GDP is relatively insensitive to the choice of conditioning variables. Note, however, that in both the regressions generating the bounds, there is convincing evidence that the residuals are not normally distributed.

This finding suggests that we should think about more robust methods than least squares. Regressions (3) and (4) report the lower and upper bounds when a selection of outliers is identified and excluded, in the manner described above. Note that the upper bound has now risen

TABLE 1
Extreme bounds on initial GDP[a,b]

	Regression				
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
Bound	Lower	Upper	Lower	Upper(1)	Upper(2)
Outliers excluded	No	No	Yes	Yes	Yes
RGDP60	−0.455	−0.345	−0.459	−0.042	−0.168
	(0.135)	(0.139)	(0.104)	(0.134)	(0.116)
Extra variables	*GDC*	*PI*	*STDC*	*GOV*	*GOV*
	X	*STPI*	*REVC*	*PI*	
	REVC			*GDC*	
Observations	86	102	83	73	84
R^2	0.56	0.48	0.71	0.79	0.77
Normality	0.04	0.00	0.83	0.23	0.98
HSW	0.90	0.14	0.77	0.87	0.72
HSF	0.82	0.45	0.05	0.05	0.75
RESET	0.67	0.74	0.48	0.36	0.20

[a] Dependent variable: growth of real per capita GDP, 1960–89.
[b] The table reports the coefficient on *RGDP60* in regressions which also include a constant, school enrolment (*SEC*), population growth (*GPOP*), average investment (*INV*), and the extra variables listed. The diagnostic tests are described in the text.

considerably. Indeed, the coefficient on initial GDP is no longer significantly different from zero even at the 50% level.

This result is surprising, to say the least. What appears to be driving it, from inspection of the various regressions that lead to the bounds, is the inclusion of the variable *GDC*. This measures the average growth rate of domestic credit over the sample period. Given that the dependent variable is the growth rate, *GDC* is almost certainly endogenous, and basing an upper bound on a regression which includes it should perhaps trouble us. When the bounds are based on a set of regressions that does not include *GDC*, they are considerably narrower. Regression (5) shows the new upper bound on the initial GDP coefficient, which appears to be more in line with what one might expect.

Overall, the results suggest that initial GDP is indeed a robust determinant of growth (see also Ley and Steel, 1999). Even so, it is clear that allowing for outliers increases the degree of uncertainty surrounding the coefficient on initial income, and hence the degree of uncertainty surrounding cross-section estimates of the rate of conditional convergence. Further discussion of the latter issue can be found in Temple (1998, 1999) and the references therein.

I now turn to the effect of real exchange rate distortions, as captured by the variable *RERDB* constructed by Dollar (1992). Regressions (1) and (2) in Table 2 report my attempt to replicate the results of Levine and Renelt; again the results are close to theirs. Note that even for the upper bound, the effect of *RERDB* is significant at the 10% level. Also note the results of the diagnostic tests. The residuals appear to be non-normal in the regression generating the lower bound. As for the regression generating the upper bound, its specification is called into question by the results of a Ramsey RESET test.

Although Levine and Renelt classify this variable as fragile, it appears to hold up quite well in some respects, even though investment is included in the regressions. In a set of regressions excluding outliers and the suspect *GDC* variable, *RERDB* is always significant at the 5% level. Regressions (3) and (4) in the table show bounds on the coefficient estimate that may be more reliable than those of Levine and Renelt. Again, though, it should be noted that there is some evidence that calls into question the specification of this regression. Even these bounds may be too wide.

In summary, it appears to be useful to supplement any EBA with diagnostic tests, at least for the regressions generating the upper and lower bounds. Presenting results for these regressions can draw attention to particularly problematic variables. As one might expect from previous

TABLE 2
Extreme bounds on real exchange rate distortions

	Regression			
	(1)	*(2)*	*(3)*	*(4)*
Bound	Lower	Upper	Lower	Upper
Outliers excluded	No	No	Yes	Yes
RERDB	−0.0196	−0.0115	−0.0240	−0.0136
	(0.006)	(0.006)	(0.004)	(0.004)
Extra variables	*REVC*	*GDC*	*X*	*STDC*
	X	*GOV*	*STDC*	*STPI*
		PI	*PI*	*REVC*
Observations	63	59	49	46
R^2	0.54	0.57	0.79	0.74
Normality	0.05	0.23	0.73	0.28
HSW	0.99	0.93	0.84	0.70
HSF	1.00	0.78	n.a.	n.a.
RESET	0.21	0.02	0.12	0.65

[a] Dependent variable: growth of real per capita GDP, 1974–89.
[b] The table reports the coefficient on *RERDB* in regressions which also include a constant, school enrolment (*SED*), population growth (*GPOP*), average investment (*INV*), initial GDP, and the extra variables listed. The diagnostic tests are described in the text; the entries ‘n.a.’, not available, indicate that there are too few observations to calculate the relevant diagnostic.

papers, it also appears that these kind of results can be sensitive to outliers. When these various factors are taken into account, the findings of an extreme bounds analysis are potentially very illuminating.

At this point, it is worth noting that over a third of the upper and lower bounds reported by Levine and Renelt include the variable *GDC*, one which is almost certainly endogenous. This suggests that some of the bounds should be approached with a certain degree of mistrust. More generally, it reminds us that whatever the chosen approach to model uncertainty, we have to be very careful about selecting the class of models we consider.

VI. CONCLUSIONS

This paper has drawn attention to three econometric problems sometimes present in applied work, and that are likely to be particularly crucial in cross-country growth regressions. Some discussion has been given to model uncertainty, parameter heterogeneity, and outliers. Drawing heavily on the existing statistics literature, the paper suggests ways in which these problems can be addressed in a tractable way by applied researchers.

One of the main arguments is that extreme bounds analysis is a useful way of communicating any uncertainty surrounding the choice of model, and hence uncertainty surrounding parameter estimates and standard errors. The form of presentation of the EBA is important. The more information that can be presented about each regression, and particularly the ones generating the bounds, the better. Many of the traditional criticisms of EBA can be addressed simply through a relatively careful presentation of the results.

Given that an EBA is essentially a means of assessing the degree of support for various relationships, I have also advocated combining its use with some investigation of the outlier problem. The paper sets out the reasons for using robust estimation methods, and indicates one way in which these methods can be applied in the course of extreme bounds analysis. These methods have wider applicability and should surely play a greater role in the growth literature than they have done to this point.

Using these ideas, I have carried out an EBA for two variables considered by Levine and Renelt. In one case, that of initial GDP, the approach suggested here leads to wider bounds. In the other case, that of an index of real exchange rate distortions, the bounds are narrowed sufficiently that classifying the relationship as ‘fragile’ may no longer be justified.

One final observation may be of interest. An underlying aim of this paper has been to contribute to making our findings in the growth literature more robust, but any test of robustness only tells us so much.

The litmus test for the cross-country growth literature will come when we find out how useful our current models are in predicting the variation in growth rates, not for existing data, but for periods beyond the usual samples. From this perspective, we must wait until around the year 2005 to see whether empirical models originally estimated using data from 1960–85 perform well over the 1985–2005 period. It is, to say the least, an intriguing prospect.

REFERENCES

Appa, G. M. and Land, A. H. (1993). 'Comment on Hettmansperger and Sheather', *American Statistician*, vol. 47, pp. 160–2.

Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Clarendon Press, Oxford.

Atkinson, A. C. (1986). 'Masking unmasked', *Biometrika*, vol. 73, pp. 533–41.

Atkinson, A. C. (1994). 'Fast very robust methods for the detection of outliers', *Journal of the American Statistical Association*, vol. 89, pp. 1329–39.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*. John Wiley, New York.

Bollen, W. A. and Jackman, R. W. (1990). 'Regression diagnostics: an expository treatment of outliers and influential cases', in: Fox, J. and Long, J. S. (eds), *Modern Methods of Data Analysis*. Sage, Newbury Park, California.

Box, G. E. P. (1980). 'Sampling and Bayes' inference in scientific modelling and robustness', *Journal of the Royal Statistical Society, Series A*, vol. 143, pp. 383–404.

Breusch, T. S. (1985). 'Simplified extreme bounds', in: Granger, C. W. J. (ed.), *Modelling Economic Series: Readings in Econometric Methodology*. Clarendon Press, Oxford (1990).

Chang, W. H., McKean, J. W., Naranjo, J. D. and Sheather, S. J. (1999). 'High-breakdown rank regression', *Journal of the American Statistical Association*, vol. 94, pp. 205–19.

Chatfield, C. (1995). 'Model uncertainty, data mining and statistical inference', *Journal of the Royal Statistical Society, Series A*, vol. 158, pp. 419–44.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley, New York.

Chesher, A. D. (1984). 'Testing for neglected heterogeneity', *Econometrica*, vol. 52, pp. 865–72.

Davies, L. and Gather, U. (1993). 'The identification of multiple outliers', *Journal of the American Statistical Association*, vol. 88, pp. 782–92.

Dollar, D. (1992). 'Outward-oriented developing countries really do grow more rapidly: evidence from 95 LDCs, 1976–85', *Economic Development and Cultural Change*, vol. 40, pp. 523–44.

Draper, D. (1995). 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society, Series B*, vol. 57, pp. 45–97.

Durlauf, S. and Johnson, P. A. (1995). 'Multiple regimes and cross-country growth behavior', *Journal of Applied Econometrics*, vol. 10, pp. 365–84.

Dutta, J. and Leon, H. L. (1991). 'Testing for heterogeneous parameters in least-squares approximations', *Review of Economic Studies*, vol. 58, pp. 299–320.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

Ferretti, N., Kelmanksy, D., Yohai, V. J. and Zamar, R. H. (1999). 'A class of locally and globally robust regression estimates', *Journal of the American Statistical Association*, vol. 94, pp. 174–88.

Fung, W. (1993). 'Unmasking outliers and leverage points: a confirmation', *Journal of the American Statistical Association*, vol. 88, pp. 515–19.

Geweke, J. (1999). 'Using simulation methods for Bayesian econometric models: inference, development, and communication', *Econometric Reviews*, vol. 18, pp. 1–73.

Granger, C. W. J. (ed.) (1990). *Modelling Economic Series: Readings in Econometric Methodology*. Clarendon Press, Oxford.

Granger, C. W. J. and Uhlig, H. (1990). 'Reasonable extreme-bounds analysis', *Journal of Econometrics*, vol. 44, pp. 159–70.

Granger, C. W. J. and Uhlig, H. (1992). 'Erratum: reasonable extreme-bounds analysis', *Journal of Econometrics*, vol. 51, pp. 285–86.

Granger, C. W. J., King, M. L. and White, H. (1995). 'Comments on testing economic theories and the use of model selection criteria', *Journal of Econometrics*, vol. 67, pp. 173–87.

Hadi, A. S. and Simonoff, J. S. (1993). 'Procedures for the identification of multiple outliers in linear models', *Journal of the American Statistical Assocation*, vol. 88, pp. 1264–72.

Hall, A. (1987). 'The information matrix test for the linear model', *Review of Economic Studies*, vol. 54, pp. 257–63.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley, New York.

Hendry, D. F. and Mizon, G. E. (1990). 'Procrustean economics: or stretching and squeezing data', in: Granger, C. W. J. (ed.), *Modelling Economic Series: Readings in Econometric Methodology*. Clarendon Press, Oxford.

Hettmansperger, T. P. and Sheather, S. J. (1992). 'A cautionary note on the method of least median squares', *American Statistician*, vol. 46, pp. 79–83.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). 'Bayesian model averaging: a tutorial', mimeo, Department of Statistics, Colorado State University.

Huber, P. J. (1981). *Robust Statistics*. John Wiley, New York.

Janson, M. A. (1988). 'Combining robust and traditional least squares methods: a critical evaluation', *Journal of Business and Economic Statistics*, vol. 6, pp. 415–51.

Judge, G. G., Griffiths, W. E., Carter Hill, R., Lutkepohl, H. and Lee, T. (1985). *The Theory and Practice of Econometrics*, 2nd edn. John Wiley, New York.

Kennedy, P. (1992). *A Guide to Econometrics*, 3rd edn. Basil Blackwell, Oxford.

Leamer, E. E. (1978). *Specification Searches*. John Wiley, New York.

Leamer, E. E. (1983). 'Let's take the con out of econometrics', *American Economic Review*, vol. 73, pp. 31–43. Reprinted in Granger (1990) and Leamer (1994).

Leamer, E. E. (1985). 'Sensitivity analyses would help', *American Economic Review*, vol. 75, pp. 308–13. Reprinted in Granger (1990) and Leamer (1994).

Leamer, E. E. (1994). *Sturdy Econometrics*. Edward Elgar, Aldershot.

Leamer, E. E. and Leonard, H. (1983). 'Reporting the fragility of regression estimates', *Review of Economics and Statistics*, vol. 65, pp. 306–17. Reprinted in Leamer (1994).

Lee, K., Pesaran, M. H. and Smith, R. (1997). 'Growth and convergence in a multi-country empirical stochastic Solow model', *Journal of Applied Econometrics*, vol. 12, pp. 357–92.

Levine, R. and Renelt, D. (1992). 'A sensitivity analysis of cross-country growth regressions', *American Economic Review*, vol. 82, pp. 942–63.

Ley, E. and Steel, M. F. J. (1999). 'We have just averaged over two trillion cross-country growth regressions', mimeo, Department of Economics, University of Edinburgh.

Magee, L. (1990). 'The asymptotic variance of extreme bounds', *Review of Economics and Statistics*, vol. 72, pp. 182–4.

McAleer, M. (1994). 'Sherlock Holmes and the search for truth: a diagnostic tale', *Journal of Economic Surveys*, vol. 8, pp. 317–70.

McAleer, M. and Veall, M. R. (1989). 'How fragile are fragile inferences? A re-evaluation of the deterrent effect of capital punishment', *Review of Economics and Statistics*, vol. 71, pp. 99–106.

McAleer, M., Pagan, A. R. and Volker, P. A. (1985). 'What will take the con out of econometrics?' *American Economic Review*, vol. 75, pp. 293–307. Reprinted in Granger (1990).

McCloskey, D. N. and Ziliak, S. T. (1996). 'The standard error of regressions', *Journal of Economic Literature*, vol. 34, pp. 97–114.

McKean, J. W., Sheather, S. J. and Hettmansperger, T. P. (1993). 'The use and interpretation of residuals based on robust estimation', *Journal of the American Statistical Association*, vol. 88, pp. 1254–63.

Pagan, A. R. (1987). 'Three econometric methodologies: a critical appraisal', *Journal of Economic Surveys*, vol. 1, pp. 3–24. Reprinted in Granger (1990).

Pena, D. and Yohai, V. (1999). 'A fast procedure for outlier diagnostics in large regression problems', *Journal of the American Statistical Association*, vol. 94, pp. 434–45.

Pesaran, M. H. and Smith, R. (1995). 'Estimating long-run relationships from dynamic heterogeneous panels', *Journal of Econometrics*, vol. 68, pp. 79–113.

Raftery, A. E. (1995). 'Bayesian model selection in social research', *Sociological Methodology*, vol. 25, pp. 111–63.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association*, vol. 92, pp. 179–91.

Rousseeuw, P. J. (1984). 'Least median of squares regression', *Journal of the American Statistical Association*, vol. 79, pp. 871–80.

Rousseeuw, P. J. (1993). 'Comment on Hettmansperger and Sheather', *American Statistician*, vol. 47, pp. 162–3.

Rousseeuw, P. J. and Hubert, M. (1999). 'Regression depth', *Journal of the American Statistical Association*, vol. 94, pp. 388–402.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley, New York.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). 'Unmasking multivariate outliers and leverage points', *Journal of the American Statistical Association*, vol. 85, pp. 633–9.

Rousseeuw, P. J., Van Aelst, S. and Hubert, M. (1999). 'Rejoinder', *Journal of the American Statistical Association*, vol. 94, pp. 419–33.

Ruppert, D. and Simpson, D. G. (1990). 'Comment on Rousseeuw and Van Zomeren', *Journal of the American Statistical Association*, vol. 85, pp. 644–6.

Sala-i-Martin, X. X. (1997). 'I just ran two million regressions', *American Economic Review*, vol. 87, pp. 178–183.

Stromberg, A. J. (1993). 'Comment on Hettmansperger and Sheather', *American Statistician*, vol. 47, pp. 87–8.

Swartz, S. and Welsch, R. E. (1986). 'Applications of bounded-influence and diagnostic methods in energy modelling', in: Belsley, D. A. and Kuh, E. (eds), *Model Reliability*. MIT Press, Cambridge, MA.

Temple, J. R. W. (1998). 'Robustness tests of the augmented Solow model', *Journal of Applied Econometrics*, vol. 13, pp. 361–75.

Temple, J. R. W. (1999). 'The new growth evidence', *Journal of Economic Literature*, vol. 37, pp. 112–56.

Temple, J. R. W. and Johnson, P. A. (1998). 'Social capability and economic growth', *Quarterly Journal of Economics*, vol. 113, pp. 965–90.

White, H. (1980). 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica*, vol. 48, pp. 817–38.

COMPUTING APPENDIX

Diagnostic tests were calculated using the software package PcGive. The EBA was carried out using computer programs written in the statistics language S-Plus. The least trimmed squares (LTS) estimator is difficult to implement in a way that is exact; S-Plus uses a genetic algorithm which approximates the estimator. This approximation is thought to be reasonably accurate, at least when fewer than 20 parameters are being estimated. In any case, the LTS estimates are used only to identify outliers, and the sets of observations identified as outliers (those with high residuals) should not vary much across different LTS estimates. Finally, note that the LTS estimator used in this paper, due to Rousseeuw (1984), should not be confused with the very different 'trimmed least squares' estimator sometimes referred to elsewhere. More information on the S-Plus software can be found at: http://www.mathsoft.com/splus/.