We Ran One Regression

David F. Hendry and Hans-Martin Krolzig*
Department of Economics, Oxford University.

March 10, 2004

Abstract

The recent controversy over model selection in the context of 'growth regressions' has led to some remarkably numerous 'estimation' strategies, including 4 million regressions by Sala-i-Martin (1997b). Only one regression is really needed, namely the general unrestricted model, appropriately reduced to a parsimonious encompassing congruent representation. Such an outcome was achieved in one run on *PcGets*, within 15 minutes of receiving from Professor Ley the data set in Fernández *et al* (2001). We reproduce that equation, and corroborate the findings in Hoover and Perez (2004), who also adopt an automatic general-to-simple approach.

1 Introduction

The literature on 'growth regressions' has growth almost as fast as the number of estimates reported by various authors: see *inter alia*, Barro and Sala-i-Martin (1995), Levine and Renelt (1992), Sala-i-Martin (1997a, 1997b), Temple (1999), Fernández, Ley and Steel (2001) and Hoover and Perez (2004). Literally millions of regressions have been estimated, and some readers may deem the entire exercise pure 'data mining', where zero reliability can be attached to anything anyone reports. We strongly disagree with such a view, and demonstrate that it has essentially no substance as an issue of model selection. En route, we explain why so many models may have been estimated, how pointless that was, and why it does not impugn inference, although it most certainly wastes time and resources of both investigators and readers. Our reported results, as the title suggests, are based on only one regression, namely the general unrestricted model (GUM) appropriately reduced to a parsimonious encompassing congruent representation. That can be achieved by one run on *PcGets*, and in fact, was so achieved when Hendry initially received the Fernández *et al.* (2001) data set in ASCII format by email from Professor Ley, and returned the results to him in under 15 minutes, when both were visiting the IMF in August 2000. We reproduce that equation below,[1] and relate it to other recent contenders in this arena. Specifically, we precisely replicate the selection in Hoover and Perez (2004), despite using a different algorithm, albeit in the same class of general-to-simple (*Gets*) approaches based on their multi-path search proposal in Hoover and Perez (1999), a different method of handling the five imputed data sets, and a different baseline significance level.

By arguing that there is basically no issue of 'data mining' from multiple-model estimation, and that model selection is almost innocuous here, we do not thereby affirm any substance to the results

*Financial support from the ESRC under grant L138251009 is gratefully acknowledged. We are indebted to Eduardo Ley and Kevin Hoover and Stephen Perez for their respective data sets, and to Kevin Hoover for helpful comments on an earlier draft.

[1] In fact, we have substantially improved *PcGets* in the meantime, so it is not precisely the originally selected model, but is close: see Hendry and Krolzig (2003b, 2003c).

obtained. The validity of a selected model depends primarily on that of the GUM as an approximation to the data generation process (DGP), which in turn involves key considerations of the accuracy of the measurements of the data series; their conceptual adequacy for the underlying causal effects; the completeness of the information (both variables and observations); the homogeneity of the sample; the independence assumptions justifying regression; the weak exogeneity of the regressors (or instruments); and the constancy of the parameters across the observations. All of these are open to legitimate doubt in this context. Nevertheless, one aspect that seems to have been a focus of debate, namely data mining, is not of great concern, and we consider it essential to clarify that issue, using this data set as an illustration. Importantly, since selection can be resolved in seconds given the GUM, we also argue that adopting automatic model selection devices such as *PcGets* (or the many related alternatives now available, including, but not restricted to, Phillips, 1994, 1995, 1996; White, 2000; Perez-Amaral, Gallo and White, 2003; Kurcewicz and Mycielski, 2003; and Hoover and Perez, 2004), frees investigators to allocate much more of their time and effort to improving the theory, data measurement and econometric specification underpinning the GUM, which in turn should improve substantive inferences in all areas of econometrics. An additional bonus is reducing the subjectivity of the selection, as is manifest from our replicating the Hoover and Perez (2004) selection: another of Keynes (1940)'s sarcastic criticisms of Tinbergen (1940) is repudiated.

2 Some model selection principles

Consider a databank containing $n+1$ variables, which defines the universe of available measures. An agnostic investigator interested in modelling one variable, say y_t decides to include all the remaining n variables $\mathbf{z_t}$ as regressors. Let us abstract from the myriad of problems noted in the introduction, to focus on the impact of selection *per se*.

First, what are the likely properties of inference in the GUM:

$$y_t = \boldsymbol{\beta}' \mathbf{z_t} + \mathbf{u_t} \tag{1}$$

where $n \leq T$, the available sample size?[2] Here, the DGP is assumed to be nested in the GUM, with $\mathsf{E}[\mathbf{z_t u_t}] = \mathbf{0}$ and $u_t \sim \mathsf{IN}[0, \sigma^2]$. Then the least squares estimates $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and $\widehat{\sigma}^2$ of σ^2 are:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{T} \mathbf{z_t z_t'} \right)^{-1} \sum_{t=1}^{T} \mathbf{z_t y_t} \text{ and } \widehat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{\mathbf{1}}{\mathbf{T-n}} \sum_{\mathbf{t=1}}^{\mathbf{T}} \left(\mathbf{y_t} - \widehat{\boldsymbol{\beta}}' \mathbf{z_t} \right)^{\mathbf{2}} \tag{2}$$

which are unbiased estimators of their respective parameters, and have the independent distributions:

$$\sqrt{T} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \sim \mathsf{N} \left[\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}} \left(\sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{z_t z_t'} \right)^{-\mathbf{1}} \right] \tag{3}$$

and:

$$\frac{\widehat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{T-n}}{T-n}. \tag{4}$$

Inference in (1) on the basis of (3) and (4) is valid, but may be inefficient if many of the $\boldsymbol{\beta}$ coefficients are zero in the population. Below, we let $\widetilde{\sigma}^2$ denote an estimator of σ^2 not corrected for degrees of freedom:

$$\widetilde{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left(y_t - \widehat{\boldsymbol{\beta}}' \mathbf{z_t} \right)^2. \tag{5}$$

[2] We later discuss the case $n > T$, but $k \leq T$ where k is defined in equation (6) below.

Secondly, what are the likely properties of inference in a simplified model selected from (1), say:

$$y_t = \boldsymbol{\theta}' \mathbf{x_t} + \mathbf{v_t} \tag{6}$$

where $\mathbf{x_t}$ is a k-dimensional sub-vector of $\mathbf{z_t}$ ($k \leq n$), chosen by some set of criteria? Naturally, the outcome depends on the selection rule, so we note the properties of three distinct approaches:
(a) information criteria such as those proposed by Akaike (1973) (denoted *AIC*), Hannan–Quinn (denoted *HQ*: see Hannan and Quinn, 1979), and Schwarz (*BIC*: see Schwarz, 1978);
(b) sifting through possible models for one that satisfies prior beliefs; and
(c) applying a procedure like *PcGets*.
The theory behind additional approaches such as extreme bounds (see Leamer, 1983, and Leamer and Leonard, 1983), and model averaging as advocated by Fernández *et al.* (2001), are not considered here. The former was criticized in McAleer, Pagan and Volker (1985), Breusch (1990) and Hendry and Mizon (1990), and examined by Temple (1999) and Hoover and Perez (2004); and the latter is susceptible to counter-examples even for forecasting, as in Hendry (2004) (for weights based on *BIC*), although we compare our empirical findings to theirs in section 3.2.

2.1 Model selection based on information criteria

First, if a model selection criterion, such as the smallest value of *AIC*, *BIC*, or *HQ*, is used, then either an asymptotically efficient (*AIC*) or consistent (*BIC*, *HQ*) selection is made. Each of these criteria penalizes the log-likelihood by $2nf\left(T\right)/T$ for n parameters and a sample size of T, where:

$$\mathit{AIC}_n = \ln \widetilde{\sigma}^2 + \frac{2n}{T};\ \mathit{BIC}_n = \ln \widetilde{\sigma}^2 + \frac{n \ln\left(T\right)}{T};\ \mathit{HQ}_n = \ln \widetilde{\sigma}^2 + \frac{2n \ln\left[\ln\left(T\right)\right]}{T}. \tag{7}$$

The penalty function is intended to reflect the 'cost' of over-parameterization. However, as shown in Campos, Hendry and Krolzig (2003), one can improve the performance of such information criteria markedly in the present context of n relatively large compared to a small T by adopting some of the procedures in *PcGets* (also see Hurvich and Tsai, 1989, who address selection when n is close to T).

Consider *BIC*: there are $2^n = M$ possible models. Let $n = 41$ as in Sala-i-Martin (1997a) (or indeed, Hoover and Perez, 1999) when $T = 72$. Then $M = 2^{41} \simeq 2.2 \times 10^{12}$ which is roughly two trillion possible models. Hoover and Perez (2004) take $n = 62$ for $T = 138$, so now we have $M = 4.6 \times 10^{18}$ (a billion billion or a million trillion) models. Is statistical inference impugned by the action of estimating all these M models? The penalty function in (7) is intended to address that issue, albeit that the three shown differ in their severity: as n falls from 41 to 1 when $T = 72$ (say), they drop from 1.15 to 0.03 for *AIC*; 2.4 to 0.06 for *BIC*; and 1.7 to 0.04 for *HQ*. To select the 'optimum' specification, all M models need to be considered, yet the result is either asymptotically efficient or consistent. However, the sample size seems too small here to justify a selected model on such grounds alone. In any case, the computational cost would be prohibitive: the second case would cost $10m even at a billion models per US cent.

2.2 Sifting through all possible models

Consider next an investigator who searches across all M models for one which confirms some theoretical views or political prejudices, rather than minimises an information criterion as in section 2.1. Assuming there exists a 'best' member on such criteria, and that such a model is found and reported, what status should it enjoy? An excellent answer was provided in Gilbert (1986) who distinguished 'weak data mining', whereby corroboration was sought for prior beliefs without seriously harming the resulting model,

from 'strong data mining' in which conflicting evidence was camouflaged or not reported. Encompassing tests against rival models can reveal the latter (also see Hendry, 1995, Ch. 15). If the reported model is not parsimoniously encompassed by any contender, it holds a place in the set of non-rejected models.

Should one worry when almost every possible combination of hypotheses has been tested in every possible model? After all, there are about $n/2$ coefficients in the average model, delivering approximately $N = n2^n/2$ tests in total. Subject to the fact that there is not a perfect match between the confidence regions of scalar and vector tests (see e.g., Scheffé, 1953, and Savin, 1980), so some extra outcomes may be squeezed into the 'corners' of such regions, then for a given significance level α using a t-test based strategy for retaining/deleting variables, αn regressors will be retained by chance (i.e., be adventitiously significant) in the final model. If $\alpha = 0.025$ (say), then for $n = 41$ that is *one* variable; for $n = 62$ it is a more awkward 1.5 (at 5%, however, it would be 3). Finding 17 or 20 variables significant out of 41 cannot be explained by simply searching across all the possible models for one that is 'preferred'. However, computing N such tests wastes even more resources, especially if the final selection is easily encompassed.

2.3 Gets-based selection

Finally, we note the *PcGets* approach, where its properties, and details of the algorithm, are described in Hendry and Krolzig (2001, 2003b, 2003c). If there are n regressors in (1), for a baseline critical value c_α for a t-test, αn regressors are retained by chance. Selecting $\alpha = 0.025$, that would lead to one 'spuriously significant' variable in the GUM and selected models shown below for $n = 41$, so again almost none of the observed significance can be ascribed to chance. Thus, there is no problem of data mining by this third route.

If all the regressors were mutually orthogonal, then the ordered squared t-statistics from the GUM, say $\mathsf{t}^2_{(1)} \geq \mathsf{t}^2_{(2)} \geq \cdots \geq \mathsf{t}^2_{(n-1)} \geq \mathsf{t}^2_{(n)}$, would suffice for the selection decision, where $\mathsf{t}^2_{(k)} \geq \mathsf{c}_\alpha$ but $\mathsf{t}^2_{(k+1)} < \mathsf{c}_\alpha$. Thus, precisely one regression would be needed: the multi-path search introduced by Hoover and Perez (1999) is essentially required to highlight what the 'underlying' t-values really are.

3 Empirical growth models

3.1 Hoover–Perez

A complication here is the multiple imputation of the missing data discussed in Hoover and Perez (2004), which requires either a 'mixed' selection strategy, such as they adopted, or a 'stacking' approach. We experimented with both: see their web site http://www.econ.ucdavis.edu/faculty/kdhoover/index.html.

First, for each of the five versions of the data set, we applied *PcGets* to select a model from the linear GUM of 62 candidate variables. This delivered 5 distinct, but highly overlapping, selections, from which we then formed the union of their 24 candidate variables. To select a final contender therefrom, we stacked the five data sets as one, estimated the union model, and applied *PcGets* (and *PcGive*: see Hendry and Doornik, 2001) to select the final choice. The logic of this is as follows.

Consider stacking the same data twice, for a basic regression equation of the form $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{u}$ with k regressors:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y} \text{ and } \mathsf{V}\left[\widehat{\boldsymbol{\beta}}\right] = \widehat{\sigma}^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1},$$

so that for a double stacked illustration:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{Z} \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \mathbf{u} \\ \mathbf{u} \end{pmatrix}.$$

Then:

$$\widetilde{\boldsymbol{\beta}} = \left(\left[\mathbf{Z}' : \mathbf{Z}' \right] \left[\begin{array}{c} \mathbf{Z} \\ \mathbf{Z} \end{array} \right] \right)^{-1} \left(\mathbf{Z}' : \mathbf{Z}' \right) \left(\begin{array}{c} \mathbf{y} \\ \mathbf{y} \end{array} \right) = \left(2\mathbf{Z}'\mathbf{Z} \right)^{-1} 2\mathbf{Z}'\mathbf{y} \equiv \widehat{\boldsymbol{\beta}} \tag{8}$$

with:

$$\overline{\sigma}^2 = \frac{1}{2T-k} \sum_{t=1}^{2T} \left(y_t - \widehat{\boldsymbol{\beta}}' \mathbf{z_t} \right)^2$$

leading to:

$$\mathsf{V}\left[\widetilde{\boldsymbol{\beta}} \right] = \overline{\sigma}^2 \left(2\mathbf{Z}'\mathbf{Z} \right)^{-1} = \frac{\overline{\sigma}^2}{2\widehat{\sigma}^2} \mathsf{V}\left[\widehat{\boldsymbol{\beta}} \right] = \frac{\mathsf{T}-\mathsf{k}}{2\mathsf{T}-\mathsf{k}} \mathsf{V}\left[\widehat{\boldsymbol{\beta}} \right]. \tag{9}$$

Consequently, (8) shows that an identical estimate of $\boldsymbol{\beta}$ would be obtained, with a slightly downward biased estimate $\overline{\sigma}^2$ of σ^2 based on conventional formulae. Also, (9) reveals the need to re-scale standard errors by a factor $\sqrt{[(2T-k)/(T-k)]}$, and hence t-values accordingly, but all of these are trivial to correct.

When different measures of the same underlying variables are used, this device is a simple way to pool, which exploits the automatic selection capabilities of *PcGets*. Thus, on the five-stacked data set, *PcGets* was applied to the 24-variable union model, but now requiring absolute t-values in excess of 5.16 for retention (namely, the 2.5% critical value of 2.267 times the scale factor $\sqrt{[(5T-k)/(T-k)]} \simeq 2.276$ for $k = 6$). *PcGive* was used in a 'hand simplification' for the last few steps, eliminating from 16 variables *seriatim* with $|\mathsf{t}| < 5.16$, as the minimum significance levels in *PcGets* were reached for smaller critical values than 5.16 (only the single path of increasing $|\mathsf{t}|$ was explored)

Correcting the estimated standard errors of the selected model by 2.276 delivered an identical selection to Hoover and Perez (2004):

$$\begin{aligned} \widehat{\mathrm{GR}}_t \;=\; & \underset{(0.002)}{0.011} + \underset{(0.004)}{0.019}\ \mathrm{YrsOpen}_t + \underset{(0.039)}{0.106}\ \mathrm{EQINV}_t + \underset{(0.017)}{0.077}\ \mathrm{CONFUC}_t \\ & \underset{(0.005)}{0.013}\ \mathrm{REVCOUP}_t - \underset{(0.005)}{0.012}\ \mathrm{PROT}_t \end{aligned} \tag{10}$$

$$\mathsf{R}^2 \;=\; 0.439 \quad \widehat{\sigma} = 0.0134 \quad \textit{BIC} = -8.57$$

Parameter constancy and normality were accepted, but there was considerable residual hereoscedasticity, possibly due to the stacking. The outcome in (10) is identical to that from the *Gets* procedure in Hoover and Perez (2004), despite using stacking, rather than averaging, and 2.5% significance, rather than their 5% rule (which could retain some other effects in the stacked approach). However, the calculated uncertainties in (10) differ slightly from those reported by Hoover–Perez, probably due to the different methods for handling the five data sets.We then repeated the selection using the stacked data throughout, applying *PcGets* first then *PcGive* for a 'hand simplification', and obtained precisely the same specification as (10).

The probability p_0 that no coefficients are significant by chance under the null given 62 orthogonal candidate variables for the criterion $|\mathsf{t}| > 2.267$ is:

$$\mathsf{p}_0 = (1 - 0.025)^{62} \simeq 0.21.$$

Thus, it cannot be precluded that all of the variables in (10) are 'genuinely' significant, in the sense of falling in that 20% of draws: this, of course, is far from sufficient to establish that they are the correct driving forces in a 'causal' sense. More generally, letting:

$$\mathsf{p}_\mathsf{j} = \frac{\mathsf{n}!}{\mathsf{j}!\,(\mathsf{n}-\mathsf{j})!} \alpha^\mathsf{j} (1-\alpha)^{\mathsf{n}-\mathsf{j}} \quad \mathsf{j} = 0, \ldots, \mathsf{n} \tag{11}$$

the next few probabilities are $\mathsf{p}_1 \simeq 0.33$, $\mathsf{p}_2 \simeq 0.26$, $\mathsf{p}_3 \simeq 0.13$, $\mathsf{p}_4 \simeq 0.05$, $\mathsf{p}_5 \simeq 0.01$ after which the probabilitiess are negligible. Although such calculations suggest a high probability of several spurious variables, the small size of the final model is really the surprise. Moreover, if $\alpha = 0.01$ is used instead, so $c_\alpha = 2.615$ then REVCOUP and PROT are eliminated, but the first row of (10) remains: now $\mathsf{p}_0 \simeq 0.54$ and $\mathsf{p}_1 \simeq 0.34$ with all other values negligible, so the first row is quite likely to be substantive—and would remain so even if 4.6×10^{18} models had been estimated.

From first downloading the Hoover and Perez (2004) data, through seeking clarification from the authors about the multiple imputation, to completing the study took a little over 2 hours of human input, including stacking the data sets and computing the corrected statistics for (10).

However, the imputation process certainly induces measurement errors in the variables, possibly correlated across measures and variables. This would generally act to bias downward the estimated coefficients, upward bias $\widehat{\sigma}$, and so probably downward bias the t-ratios, leading to underselection. Conversely, endogeneity of the variables would act in the opposite direction. Similarly, many of the variables have an anticipated sign in any sensible model, so could be tested on a 1-sided significance level.

The selection was unaffected by allowing for outlier removal in the GUM and all subsequent models.

3.2 Fernández *et al*

The original motivation of our paper, however, was as a comment on Fernández *et al.* (2001), so we briefly note the comparative outcomes on their data set. The GUM regression for all 41 variables had $\widehat{\sigma} = 0.0056$ with $BIC = -8.743$, and insignificant diagnostic tests for constancy and normality (heteroscedasticity could not be calculated for $T = 72$). Using an overlapping notation with Hoover and Perez (2004), but the order in Fernández *et al.* (2001), the specific model selected by precisely the same settings for the algorithm was:

$$\begin{aligned}
\widehat{\mathrm{GR}}_t \;=\; & \underset{(0.022)}{-\,0.165}\ \mathrm{GDPSH60L}_t + \underset{(0.010)}{0.056}\ \mathrm{CONFUC}_t + \underset{(0.021)}{0.098}\ \mathrm{LIFEE060}_t + \underset{(0.039)}{0.186}\ \mathrm{EQINV}_t \\
& - \underset{(0.003)}{0.027}\ \mathrm{SAfrica}_t + \underset{(0.004)}{0.015}\ \mathrm{RuleLaw}_t + \underset{(0.013)}{0.034}\ \mathrm{Mining}_t - \underset{(0.004)}{0.016}\ \mathrm{LAAM}_t \\
& + \underset{(0.006)}{0.017}\ \mathrm{P60}_t - \underset{(0.019)}{0.111}\ \mathrm{Hindu}_t + \underset{(0.004)}{0.013}\ \mathrm{EthnoLing}_t + \underset{(0.004)}{0.015}\ \mathrm{SpanishCol}_t \qquad (12) \\
& + \underset{(0.003)}{0.011}\ \mathrm{FrenchCol}_t + \underset{(0.001)}{0.004}\ \mathrm{LabForce}_t - \underset{(0.030)}{0.144}\ \mathrm{HYR}_t + \underset{(0.002)}{0.008}\ \mathrm{BritCol}_t + \underset{(0.015)}{0.059} \\
\mathsf{R}^2 \;=\; & 0.907 \;\; \widehat{\sigma} = 0.0063 \;\; BIC = -9.389
\end{aligned}$$

Neither constancy nor normality diagnostic tests rejected. There is considerable overlap with the findings in Fernández *et al.* (2001), but we do not confirm some of their claimed more probable averaged variables. Several variables found with high significance by Sala-i-Martin (1997a) are not replicated here.

Perhaps the most salient difference with section 3.1 is the large number of variables selected for the subset of countries where all observations are available on the 41 candidate regressors. This could reflect a different source of selection bias (e.g., choice of observations), greater endogeneity in the sub-sample, or the limitations of multiple imputation.

4 Conclusion

The efficiency gains from automatic procedures for investigators wishing to undertake model selection are potentially huge. We ran one basic regression for each data set, precisely so for Fernández *et al.* (2001), though really requiring 6 regressions for Hoover and Perez (2004) due to the multiply-imputed data. This contrasts with the millions actually estimated by Sala-i-Martin (1997a, 1997b), despite which the likely number of spuriously-significant variables could be calculated as one out of 41.

The main point of our note is that the results obtained thereby are unlikely to be due to chance significance in a setting where the data generation process is a special case of the general model postulated at the outset and a *Gets* approach is adopted. However, that assumption is hard to believe here, and all the other usual issues remain applicable to the empirical modelling.

We noted above that similar considerations apply when $n > T$, even though the GUM cannot be directly estimated. All that is required is a repeated selection algorithm, not dissimilar to that used above for handling the five imputations of data: see e.g., Hendry and Krolzig (2003a, 2003d), and Hendry, Johansen and Santos (2004). Naturally, a larger n suggests a smaller α, as with information criteria, but selected to reflect an investigator's tradeoff of adventitious significance against omitting relevant variables. Thus, surprising new insights can follow from thinking based on the theory of reduction which underpins *Gets*.

An application of considerable interest here is whether country-specific dummies are required (the imputation stacking actually entails each 'impulse' having 5 values of unity). Doing so would check one aspect of the homogeneity of the sample relative to the selected model (strictly one should add the impulses to the GUM, but the computations are excessive without a special program—which still awaits development). Under the null, at 0.5%, adding 138 country dummies should produce less than one significant by chance, but could reveal important heterogeneities if they existed (the many regional dummies in Sala-i-Martin, 1997a, presumably attempt to capture such). This emphasizes the two main points of our comment, namely that repeated testing is not very harmful; and that automatic methods can eliminate what would otherwise be intolerable computational burdens.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., and Csaki, F. (eds.), *Second International Symposion on Information Theory*, pp. 267–281. Budapest: Akademia Kiado.

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

Breusch, T. S. (1990). Simplified extreme bounds. In Granger (1990), pp. 72–81.

Campos, J., Hendry, D. F., and Krolzig, H.-M. (2003). Consistent model selection by an automatic Gets approach. *Oxford Bulletin of Economics and Statistics*, **65**, 803–819.

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

Gilbert, C. L. (1986). Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics*, **48**, 283–307. Reprinted in Granger (1990).

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

Hannan, E. J., and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, **B, 41**, 190–195.

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

Hendry, D. F. (2004). Model averaging with indicator variables. Working paper, Economics Department, Oxford University.

Hendry, D. F., and Doornik, J. A. (2001). *Empirical Econometric Modelling using PcGive 10: Volume I*. London: Timberlake Consultants Press.

Hendry, D. F., Johansen, S., and Santos, C. (2004). Selecting a regression for more indicators than observations. Unpublished paper, Economics Department, University of Oxford.

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

Hendry, D. F., and Krolzig, H.-M. (2003a). Model selection with more variables than observations. Unpublished paper, Economics Department, Oxford University.

Hendry, D. F., and Krolzig, H.-M. (2003b). New developments in automatic general-to-specific modelling. In Stigum, B. P. (ed.), *Econometrics and the Philosophy of Economics*, pp. 379–419. Princeton: Princeton University Press.

Hendry, D. F., and Krolzig, H.-M. (2003c). The properties of automatic Gets modelling. Unpublished paper, Economics Department, Oxford University.

Hendry, D. F., and Krolzig, H.-M. (2003d). Resolving three 'intractable' problems using a Gets approach. Unpublished paper, Economics Department, University of Oxford.

Hendry, D. F., and Mizon, G. E. (1990). Procrustean econometrics: or stretching and squeezing data. In Granger (1990), pp. 121–136.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, **2**, 167–191.

Hoover, K. D., and Perez, S. J. (2004). Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics*, **66**, forthcoming.

Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Keynes, J. M. (1940). Statistical business-cycle research: Comment. *Economic Journal*, **50**, 154–156.

Kurcewicz, M., and Mycielski, J. (2003). A specification search algorithm for cointegrated systems. Discussion paper, Statistics Department, Warsaw University.

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

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

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

McAleer, M., Pagan, A. R., and Volker, P. A. (1985). What will take the con out of econometrics?. *American Economic Review*, **95**, 293–301. Reprinted in Granger (1990).

Perez-Amaral, T., Gallo, G. M., and White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics*, **65**, 821–838.

Phillips, P. C. B. (1994). Bayes models and forecasts of Australian macroeconomic time series. In Hargreaves, C. (ed.), *Non-stationary Time-Series Analyses and Cointegration*. Oxford: Oxford

University Press.

Phillips, P. C. B. (1995). Automated forecasts of Asia-Pacific economic activity. *Asia-Pacific Economic Review*, **1**, 92–102.

Phillips, P. C. B. (1996). Econometric model determination. *Econometrica*, **64**, 763–812.

Sala-i-Martin, X. X. (1997a). I have just run two million regressions. *American Economic Review*, **87**, 178–183.

Sala-i-Martin, X. X. (1997b). I have just run four million regressions. Unpublished typescript, Economic Department, Columbia University, New York.

Savin, E. (1980). The Bonferroni and Scheffé multiple comparison procedures. *Review of Economic Studies*, **47**, 255–273.

Scheffé, H. (1953). A method of judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Temple, J. (1999). Growth regressions and what the textbooks don't tell you. Unpublished typescript, Nuffield College, Oxford.

Tinbergen, J. (1940). *Statistical Testing of Business-Cycle Theories. Vol. I: A Method and its Application to Investment Activity*. Geneva: League of Nations. Reprinted in part in Hendry, D. F. and Morgan, M. S. (1995), *The Foundations of Econometric Analysis*. Cambridge: Cambridge University Press.

White, H. (2000). A reality check for data snooping. *Econometrica*, **68**, 1097–1126.