The Ill-Posed Problem in Growth Empirics

Jensen, Peter Sandholt; Würtz, Allan Halkjær

Publication date:
2005

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Jensen, P. S., & Würtz, A. H. (2005). *The Ill-Posed Problem in Growth Empirics*. Cph.: Department of Economics, University of Copenhagen.

CAM

**Centre for Applied
Microeconometrics**

**Department of Economics
University of Copenhagen**

http://www.econ.ku.dk/CAM/

The Ill-Posed Problem in Growth Empirics

Peter Sandholt Jensen, Allan H. Würtz

2005-11

**The activities of CAM are financed by a grant from
The Danish National Research Foundation**

The Ill-Posed Problem in Growth Empirics

Peter Sandholt Jensen[*] Allan H. Würtz[‡]

July 12, 2005

Abstract

A problem encountered in growth empirics is that the number of explanatory variables is large compared to the number of observations. This makes it impossible to condition on all regressors when determining if a variable is important. We investigate methods used to resolve this problem: Extreme bounds, Sala-i-Martin's test, BACE, general-to-specific, minimum t-statistics, BIC and AIC. We prove that the problem in general is ill-posed and that the existing methods are inconsistent. We propose a test and apply it to determine if "good policy" increases the effectiveness of foreign aid on growth. The test rejects inference regarding good policy.

Keywords: AIC, BACE, BIC, extreme bounds, general-to-specific, ill-posed inverse problem, robustness

Jel: C12, C51, O10

[*] Department of Economics, University of Aarhus. Building 322, DK-8000 Aarhus C, Denmark. E-mail: psjensen@econ.au.dk

[†] Institute of Economics, University of Copenhagen, DK 1455 Copenhagen K, Denmark. Email: allan.wurtz@econ.ku.dk

[‡] *Acknowledgements*: We are thankful for discussions with Martin Browning, Gernot Doppelhoffer, John Geweke, Clive Granger, Niels Haldrup and Gene Savin. Comments from seminar participants at the University of Aarhus, University of Copenhagen and University of Helsinki are appreciated. Allan Würtz acknowledges the support from the Centre for Applied Microeconometrics (CAM). The activities of CAM are financed from a grant by the Danish National Research Foundation.

1 Introduction

A problem encountered in growth empirics is that the number of explanatory variables of GDP growth is large compared to the number of observations. For example, in the literature on GDP growth more than hundred variables have been suggested to explain growth, see e.g. Durlauf and Quah (1999) or Durlauf, Johnson and Temple (2004) for a list of 145 variables. Even in cases where more observations than variables are available, concerns for the precision of estimates will necessitate restricted models.

The aim of many empirical papers is to determine the importance of a variable of interest. Most authors agree that a variable is important if it belongs in a regression among other variables believed to be of importance. Naturally, importance of the variable must be qualified in the economic context, for instance, to be of a certain size. In growth empirics the regression with all variables believed to be of importance is often not feasible. The infeasibility of this regression has made researchers employ various types of model selection methods and it has inspired development of new methods.

One set of methods are Bayesian in spirit and builds on Leamer (1983). He argues that a variable is important if it is significant and has the same sign in regressions involving different subsets of the variables believed to be of importance. In that case he denotes the variable robust. The method is known as extreme bounds analysis and it was first implemented by Levine and Renelt (1992) in a growth context. Sala-i-Martin (1997a, 1997b) criticized the extreme bounds for sampling reasons since an insignificant variable is likely to be found if enough regressions are run. Sala-i-Martin's test for a robust variable is based on calculating a distribution of the parameters to the variable of interest taken over models. Sala-i-Martin, Doppelhoffer and Miller (2004) build directly on the Bayesian model averaging method from which they derived an approach called "Bayesian averaging of classical estimates". In these approaches, the robustness of the variable of interest is determined by the average of the estimated parameter values for the variable in each of the different models. Leamer's approach can also be implemented by bootstrapping the distribution of the minimal t-statistics over models, see White (2000) and Hansen (2003). For all these methods, robustness of a variable is defined for a sample, but not linked to the importance of the variable in the population. That is, when there is no sample uncertainty. In this paper, the link for all these methods is derived.

Classical variable selection and model selection methods can also be employed to determine the importance of a variable. Pure variable selection procedures determine if the variable belongs in the model or not, whereas model selection criteria as a by-product select a set of variables of which the variable of interest may be one. Criteria such as AIC and BIC can be employed to choose the best model and the importance of the variable is then determined by the chosen model. In the same vein, refined general-to-specific procedures suggested by Hoover and Perez (2004), Bleaney and Nishiyama (2002) and Hendry and Krolzig (2004) can be applied. The classical methods are also analyzed in this paper and compared with the Bayesian inspired methods.

We show with an impossibility theorem that determining the importance of a variable in the regression of all variables believed to be of importance is an ill-posed inverse problem,[1] if the number of variables is larger than the number of observations.[2] The impossibility theorem explains why none of the methods work. Still, by deriving results on the methods insight is obtained into their properties in general. Our result differs from the usual consistency results of these methods because we explicitly account for the ill-posedness of the problem. We reconcile the results with and without the ill-posed inverse problem by providing identification conditions under which the methods perform correctly. The main conclusion of all the identification results is that more information is needed to conduct correct inference on the importance of the variable of interest.

The conditions of the impossibility theorem are used to construct a new approach to determining the importance of the variable of interest. By the nature of an impossibility theorem, the approach can only work when the conditions of the impossibility theorem are not satisfied. These conditions, however, can be tested. We compare the new approach with the existing approaches in a Monte Carlo study under various assumptions which are necessary to identify importance of a variable. The results suggest that the new approach has good finite sample properties.

An important application in empirical growth is determining the effect of foreign aid on growth and the role of "good policy." Burnside and Dollar (2000) have shown that foreign aid is most effective if accompanied by good policy. Easterly, Levine and Roodman (2004) have disputed the findings of Burnside and Dollar. In view of other variables believed to influence growth, we show that the influence of good policy on the effectiveness of aid on growth cannot be determined without imposing additional information.

The outline of the paper is as follows. In section 2 we prove the impossibility theorem and provide further conditions in order to identify importance of a variable. Then in section 3, properties of the existing methods are derived. Section 4 describes a new test of importance of a variable, and in section 5 the different methods are compared in a Monte Carlo study. Section 6 considers the importance of good policy on the effectiveness of aid on growth. Section 7 concludes the paper.

[1] See Carresco, Florens and Renault (2003) for general reference on ill-posed inverse problems. Discrete cases are also denoted an ill-conditioned problems.

[2] It can also specifically be referred to as an undersized sample problem.

2 Identification

In this section we prove that determining importance of the variable of interest in general is impossible because the problem is ill-posed due to an undersized sample. As a consequence, it is necessary to search for special cases or add additional information to make the problem well-posed. Though the majority of the results in this section are known from various contexts, we want to state the results (with proofs) for a better understanding of the results in the next section on the existing methods and our suggestion on a new test in section 4.

We first prove that the importance of a regressor in general is not identified. The problem can be stated as follows. Let N be the number of observations and K the number of variables believed to be important. The problem is to determine if X_1 is important, that is, whether X_1 belongs in the regression or not. The remaining $(K-1)$ regressors are X_2,..,X_K. As it is usually done, we assume that all regressions are linear.[3] The regression with all regressors believed to be of importance is

$$E(y \mid X_1, X_2, .., X_K) = \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K. \tag{1}$$

The variable of interest, X_1, belongs in the regression if $\beta_1 \neq 0$.

The following theorem[4] shows that inference on β_1 is impossible if only using the information in (1). The reason is that the problem is ill-posed due to the undersized sample.

Theorem 1 (Impossibility) *Assume $N < K$ and β_2,.., $\beta_K \neq 0$. Partition $\{X_2, .., X_K\}$ into two sets A and B with $(N-1)$ and $(K-N-1)$ variables, respectively. If there does not exist a set A such that $E(X_j \mid X_1, A) = 0$ for all $X_j \in B$, then β_1 is not identified nor can it be bounded.*

Proof. *See appendix* ■

The theorem has a lot in common with the omitted variable bias problem in a well-posed setting but there is a difference. In the omitted variable bias problem, the coefficient to x_1 can be identified when $E(X_j \mid X_1, A)$ is known for all $X_j \in \{X_2, .., X_K\} \backslash A$ if it differs from a linear index. When the problem is ill-posed, knowing $E(X_j \mid X_1, A)$ only helps if it is equal to 0.

The theorem rules out that combinations of regressions with fewer regressors than observations can be used to infer the value of β_1. In particular, the application of the omitted variable rule, Goldberger (1998), and the Frisch-Waugh-Lovell theorem might at

[3] Alternatively, the results can be interpreted in terms of best linear predictions.

[4] In general about inference in unidentified linear regression models, see e.g. Scheffé (1959) and Rao (1973).

first look promising. To illustrate why these fail, consider the regression $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$ for $N = 2$. Then the omitted variable rule gives: $\gamma_1 = \beta_1 + \beta_2 c_2 + \beta_3 c_3$, where γ_1 is the coefficient to x_1 in the regression of y on x_1. In the case $(N \geq K)$, it is possible to infer β_2, β_3, c_2 and c_3 from short regressions and β_1 can thereby be inferred. One method is to run the auxiliary regressions $x_{2i} = \rho_2 x_{1i} + \nu_{2i}$, $x_{3i} = \rho_3 x_{1i} + \nu_{3i}$ and $y_i = \rho_y x_{1i} + \nu_{yi}$, where $\widehat{\rho}_2$ and $\widehat{\rho}_3$ are estimates on c_2 and c_3. Both of these regressions are possible in the undersized sample case. Finally, to estimate β_2 and β_3 the Frisch-Waugh-Lovell theorem says that $\widehat{\nu}_{yi} = \widehat{\beta}_2 \widehat{\nu}_{2i} + \widehat{\beta}_3 \widehat{\nu}_{3i} + \widehat{\xi}_i$. The sample size appears sufficient to run this regression. However, due to the undersized sample the original regressors are linearly dependent, for instance $x_{1\cdot} = a_2 x_{2\cdot} + a_3 x_{3\cdot}$ and by insertion it can be seen that $\widehat{\nu}_{2\cdot}$ is linearly dependent on $\widehat{\nu}_{3\cdot}$. Therefore, inversion to estimate the β's is not possible. This illustrates why an undersized sample leaves no backdoor regressions for recovering the coefficients in the long regression.

While the impossibility theorem shows that importance of a single regressor cannot be determined, it is possible to identify the importance of a set of regressors. It is not, however, possible to say which ones of them are important. Knowing that at least one regressor in a set is robust can be useful in practice. For example, if the set of regressors consists of different (known) functions of one variable, then knowing that at least one function of the variable is important means that the variable itself is important though not in what functional form. To determine if at least one variable in a set is important, the number of regressors in the set must be no less than the number of non-orthogonal regressors plus one excluded from the regressions. The next theorem states the result.

Theorem 2 (Partial identification of variables of importance) *Let r be the number of regressors in the set $R \subset \{X_1, .., X_K\}$. Partition the remaining regressors, $\{X_1,.., X_K\} \backslash R$, into C_1 with $(N - r)$ regressors and C_2 with $(K - N)$ regressors. If there exists a set C_2 such that $E(X_j \mid X_1, A) \neq 0$ for at most $(r - 1)$ of the regressors $X_j \in C_2$, then at least one of the regressors in R is important.*

Proof. See appendix. ■

Partial identification is possible because the conditions in the theorem lead to an overidentified system of equations.[5]

The impossibility theorem makes it clear that in order to make inference on β_1 in general it is necessary to include additional information than (1). This information could be in the form of an economic model, which could exclude some of the regressors. It could also be exclusion restrictions which would permit use of instrumental variables. Another possibility is to impose priors on the coefficients. Then it would be possible to estimate β_1

[5] The result is in line with results on testing in unidentified models, see Breusch (1986).

using for example some method of regularization e.g. ridge regression, see Mittelhammer, Judge and Miller (2000). Another regularization method is to condition on a subset of principal components, see e.g. Stock and Watson (2002). These methods would make estimation of β_1 possible, but the consistency of such methods depends on eventually the ill-posed inverse problem being eliminated.

One type of added information is assuming that $(K - N)$ regressors in (1) are not important. This is equivalent to assuming that at least $(K - N)$ of the β's are 0, but not which ones of them. The next theorem provides sufficient conditions for identification of the regression with only the variables of importance using a measure of model fit. The theorem has been proved in different contexts so we mainly state it for easy reference to the other results in the paper.

Theorem 3 (Identification assuming a true submodel) *Assume $A \subset \{X_1, .., X_K\}$ has at most N members and that A is the smallest subset such that*

$$E(y \mid A) = E(y \mid X_1, .., X_K).$$

Then $E(V(y \mid A)) < E(V(y \mid B))$ for any B such that $A \not\subseteq B$.

Proof. See appendix. ■

The theorem implies that a model selection criterion based on fit (in terms of $E(V(y \mid X)))$ may identify the correct model and, thus, inference on X_1 can be performed. Many of the existing methods are based on this criterion as it will be demonstrated in the next section. It is important to stress that the added information regarding variables of no importance cannot be tested.

3 Properties of existing methods

In this section we derive properties of extreme bounds, refinements of extreme bounds and other model selection methods in the context of the ill-posed inverse problem caused by an undersized sample. The impossibility theorem in Section 2 already demonstrates that no method can work correctly in a general case. The reason why we derive the results in the general case anyhow is to obtain useful insight into the properties of the different methods. We also analyze a special case where the true model is among the feasible regressions. This is the case for which consistency results of model selection methods typically are derived. It is worth stressing that our results are equally applicable in well-posed problems, for instance if the regression with all regressors believed to be of importance cannot be estimated with sufficiently high precision.

The properties of the methods are derived for the population. Doing so provides the methods with the best possible environment under which to perform well. The properties of the methods in the population are equivalent to asymptotic theory based on the number of observations $N \to \infty$ while keeping the conditioning sets of variables fixed to resemble that the problem is ill-posed in finite sample.

It is sufficient to analyze a case with four regressors to obtain the properties of the methods. Adding more variables only increase the number of different regressions but change nothing of substance. The regression with all variables believed to be of importance is:

$$E(y \mid x_1, x_2, x_3, x_4) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4. \tag{2}$$

This regression is denoted the long regression. The problem is to determine if x_1 belongs in the regression, that is, if $\beta_1 \neq 0$. Without loss of generality, assume $E(x_k) = 0$, $V(x_k) = 1$ and $Corr(x_k, x_m) = \rho_{km}$ for all k, m.

The resemblance of the ill-posed problem due to an undersized sample is done by allowing regressions with two regressors at most. These are denoted short regressions. To keep the discussion focussed on the ill-posed problem, assume there is no misspecification and that all short regressions are linear.

The coefficient to x_1 in a short regressions may be different from β_1 due to omitted variable bias. Let $[mk]$ index the regression of y on x_m and x_k, and let $\beta_{m;k} = (\beta_m, \beta_k)'$. For notational convenience $k = 0$ denotes no regressor e.g. [10] is the regression of y on x_1. Let $\gamma_1^{[1k]}$ be the coefficient on x_1 in the linear regression of y on x_1 and x_k. Then

$$\begin{aligned}
\gamma_1^{[12]} &= \beta_1 + \frac{\rho_{13} - \rho_{12}\rho_{23}}{1 - \rho_{12}^2}\beta_3 + \frac{\rho_{14} - \rho_{12}\rho_{24}}{1 - \rho_{12}^2}\beta_4 = \beta_1 + c'_{[12]}\beta_{3;4}, \\
\gamma_1^{[13]} &= \beta_1 + \frac{\rho_{12} - \rho_{13}\rho_{23}}{1 - \rho_{13}^2}\beta_2 + \frac{\rho_{14} - \rho_{13}\rho_{34}}{1 - \rho_{13}^2}\beta_4 = \beta_1 + c'_{[13]}\beta_{2;4}, \\
\gamma_1^{[14]} &= \beta_1 + \frac{\rho_{12} - \rho_{14}\rho_{24}}{1 - \rho_{14}^2}\beta_2 + \frac{\rho_{13} - \rho_{14}\rho_{34}}{1 - \rho_{14}^2}\beta_3 = \beta_1 + c'_{[14]}\beta_{2;3}, \\
\gamma_1^{[10]} &= \beta_1 + \rho_{12}\beta_2 + \rho_{13}\beta_3 + \rho_{14}\beta_4 = \beta_1 + c'_{[10]}\beta_{2;3;4}.
\end{aligned}$$

The last terms are the omitted variable biases.

We focus on two cases of general interest. The first one, denoted the generic case, is the case with virtual no restrictions on the long regression. The generic case is:

$$\begin{array}{|l|}
\hline
\text{Generic Case} \\
\hline
\beta_2 \neq 0,\ \beta_3 \neq 0,\ \beta_4 \neq 0 \\
\beta_1 \neq -c'_{[12]}\beta_{3;4}, \text{ and } c'_{[12]}\beta_{3;4} \neq 0 \\
\beta_1 \neq -c'_{[13]}\beta_{2;4}, \text{ and } c'_{[13]}\beta_{2;4} \neq 0 \\
\beta_1 \neq -c'_{[14]}\beta_{2;3}, \text{ and } c'_{[14]}\beta_{2;3} \neq 0 \\
\beta_1 \neq -c'_{[10]}\beta_{2;3;4}, \text{ and } c'_{[10]}\beta_{2;3;4} \neq 0 \\
\hline
\end{array} \tag{3}$$

The inequalities are imposed in such a way that there is an effect of all regressors and that they are correlated. The correlation implies that an omitted variable bias exists in the short regressions.

The other case, denoted the special case, is chosen so that one of the short regressions is equivalent to the long regression. Consistency of model selection methods is typically established for this case. The special case is:

Special Case
$\beta_2 \neq 0,\ \beta_3 = \beta_4 = 0$
$\beta_1 \neq -c'_{[13]}\beta_{2;4}$, and $c'_{[13]}\beta_{2;4} \neq 0$
$\beta_1 \neq -c'_{[14]}\beta_{2;3}$, and $c'_{[14]}\beta_{2;3} \neq 0$
$\beta_1 \neq -c'_{[10]}\beta_{2;3;4}$, and $c'_{[10]}\beta_{2;3;4} \neq 0$

(4)

In the special case, x_3 and x_4 are not of importance in the long regression but they are correlated with the other variables. Furthermore, the long regression is equivalent to the short regression of y on x_1 and x_2.

In the following subsections we shortly describe different methods and their properties in the generic and special case.

3.1 Extreme bounds

The extreme bounds analysis of Leamer (1983) and Levine and Renelt (1992) define the variable x_1 as robust if the estimate on the corresponding coefficients are significantly different from 0 and have the same sign in all the short regressions with x_1. Other authors have slightly different definitions of robustness, see the next subsections. All authors agree, however, that the idea of robustness is to determine if the variable is important. This leads naturally to our definition of importance of a variable in the population. Therefore, in this (and the following subsections) we derive the properties of methods that investigate robustness to see if they correctly characterize the variable as important or not.[6]

The extreme bounds have been criticized by McAleer, Pagan and Volcker (1985), Pagan (1987), Breusch (1990) and Granger and Uhlig (1990) in a well posed setting. McAleer, Pagan and Volcker (1985) derive the probability that a variable is robust. Breusch (1990) calculated the extreme bounds based on the long regression. Their results are closely related to our results below. Granger and Uhlig (1990) derived the extreme bounds over short regressions which have a reasonable fit (in terms of R^2) relative to the best and worst fitting models. McAleer(1994) reiterates the points made in McAleer et al (1985) and criticizes Levine and Renelt (1992) for not reporting diagnostic tests. Despite

[6] The results can also simply be seen as finding the asymptotic distribution for $N \to \infty$ of the various methods given the problem is ill-posed.

criticism, the extreme bounds analysis continues to enjoy popularity, see e.g. Temple (2000), de Haan and Sturm (2000) and Kalaitzidakis, Mamuneas and Stengos (2002).

The next proposition shows that the extreme bounds cannot correctly characterize x_1 as important in neither the generic nor the special case.

Proposition 4 (Extreme bounds) *In the generic case (3) the extreme bounds analysis has the following properties in the population*[7]*:*

	Truth: $\beta_1 = 0$
Correct	$sgn(c'_{[12]}\beta_{3;4}) \neq sgn(c'_{[13]}\beta_{2;4})$ *or* $sgn(c'_{[12]}\beta_{3;4}) \neq sgn(c'_{[14]}\beta_{2;3})$
Incorrect	$sgn(c'_{[12]}\beta_{3;4}) = sgn(c'_{[13]}\beta_{2;4}) = sgn(c'_{[14]}\beta_{2;3})$

	Truth: $\beta_1 \neq 0$
Correct	$\beta_1 < -\max(c'_{[12]}\beta_{3;4}, c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3})$ *or* $\beta_1 > -\min(c'_{[12]}\beta_{3;4}, c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3})$
Incorrect	$-\max(c'_{[12]}\beta_{3;4}, c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3}) < \beta_1 < -\min(c'_{[12]}\beta_{3;4}, c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3})$

In the special case (4) the properties are:

	Truth: $\beta_1 = 0$
Correct	*all cases*

	Truth: $\beta_1 \neq 0$
Correct	$\beta_1 < -\max(0, c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3})$ *or* $\beta_1 > -\min(0, c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3})$
Incorrect	$-\max(c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3}) < \beta_1 < -\min(c'_{[13]}\beta_{2;4}, c'_{[14]}\beta_{2;3})$

Proof. *See appendix.* ■

The proposition shows that the extreme bounds criterion is not a consistent procedure in determining if a regressor is important in the long regression. In particular, if $sgn(c_{[12]}\beta_{3;4}) \cdot sgn(c_{[13]}\beta_{2;4}) = 1$, then the extreme bound analysis will give an incorrect result when x_1 does not belong in the long regression, but it will also give an incorrect result when x_1 belongs in the regression for a range of values of $\beta_1(\neq 0)$. In practice, there is nothing peculiar about a case where $sgn(c_{[12]}\beta_{3;4}) \cdot sgn(c_{[13]}\beta_{2;4}) = 1$. It is all a question of the correlation between the regressors and the true values of the parameters.

[7] The $sgn()$ function is defined as $sgn(z) = \begin{cases} -1 & if \quad z < 0 \\ 0 & if \quad z = 0 \\ 1 & if \quad z > 0 \end{cases}$

The extreme bounds can be modified so that it is an almost everywhere consistent procedure in the special case. The modification is a change of the decision rule to accept that x_1 is important when the estimate of β_1 is significantly different from 0 in all short regressions. Compared to extreme bounds, a (significant) sign shift does not disqualify a variable from being important.

3.2 Sala-i-Martin's test

Sala-i-Martin (1997a, 1997b) motivates his approach as an alternative to the extreme bounds where sampling uncertainty is taken more into account. He considers a set-up in which all the short regressions, m, have the same number of explanatory variables and always include the variable of interest. In his general setup, where the estimates $(\widehat{\gamma}_1)$ across short regressions are not assumed to be normal, he suggests the statistic.[8]

$$CDF(0) = \sum_{i=1}^{m} w_i CDF_i(0),$$

where w_i is the weight of short regression i and $CDF_i\,(0) = Max(\Phi(\widehat{\gamma}_1/\widehat{\sigma}_{\widehat{\gamma}_1}), 1-\Phi(\widehat{\gamma}_1/\widehat{\sigma}_{\widehat{\gamma}_1}))$ is the largest of the areas to the left or to the right of zero in a normal distribution with mean equal to the OLS estimator $(\widehat{\gamma}_1)$ and variance equal to the variance of the OLS estimator $(\widehat{\sigma}^2_{\widehat{\gamma}_1})$. A variable is robust (denoted important) if $CDF\,(0)$ is greater than 0.95. Sala-i-Martin assumes conditional normality of y in all the short regressions. Then, the weight of model j is defined as:

$$w_j = \frac{SSE_j^{-N/2}}{\sum_{i=1}^{m} SSE_i^{-N/2}},$$

where SSE_j is the sum of squared errors in model j.

The next proposition shows that the test always characterizes the variable of interest as important in the generic case:

Proposition 5 (Sala-i-Martin's test) *In the generic case* (*3*) *Sala-i-Martin's test has the following properties in the population*:

	Truth: $\beta_1 = 0$	*Truth:* $\beta_1 \neq 0$
Correct	*none*	*all cases*
Incorrect	*all cases*	*none*

In the special case the test correctly determines the importance of the regressor.

[8] He also considers versions based on the average beta and average variance. In the general case they have similar properties as the version analyzed here. The methodology has recently been applied by Sturm and de Haan(forthcoming).

Proof. See appendix. ■

In the special case in which one of the short regressions is equivalent to the long regression, the test works because the correct specification is given a weight equal to one because this regression has the best fit in the sense of minimizing $E(V(y \mid x_m, x_k))$, see the proof. Hence, the working of the test in the special case relies on the identification result in Theorem 3.

Finally, note that the test has properties similar to the modified extreme bounds test defined in the end of the section 3.2.

3.3 BACE

The idea of Bayesian model averaging is implemented in a simplified version by Sala-i-Martin, Doppelhoffer and Miller (2004).[9] They call their version Bayesian Averaging of Classical Estimates (BACE).

All short regressions are included in the averaging, also the ones without the variable of interest. Let C^* be the total number of short regressions. The posterior probability of the j'th short regression, M_j, is:

$$P(M_j \mid y) = \frac{P(M_j)\, N^{-k_j/2} SSE_j^{-N/2}}{\sum\limits_{i=1}^{C^*} P(M_i)\, N^{-k_i/2} SSE_i^{-N/2}}, \tag{5}$$

where $P(M_i)$ is the prior probability of model i.[10] The (classical) estimate, $\widehat{\beta}_1^{SDM}$, of β_1 is the weighted average of the estimates from each model by the model posterior probabilities:

$$\widehat{\beta}_1^{SDM} = \sum_{i=1}^{C^*} \widehat{\beta}_1^i P(M_i \mid y),$$

where $\widehat{\beta}_1^i$ is the estimated value of β_1 in short regression i.

In the next proposition it is shown that BACE selects models based on fit rather than the importance of the variable of interest.

Proposition 6 (BACE) *Define* $\sigma^2_{[mk]} = E(V(y \mid x_m, x_k))$ *and* $\sigma_{\min} = \min\limits_{i,j}(\sigma_{[ij]})$. *Let* $S_1 = \{k \in \{0, 2, .., K\} \mid \sigma_{[1k]} = \sigma_{\min}\}$ *be the set of indices for regressions that include* x_1 *and achieve the smallest variance.*

[9] See also Fernandez, Ley and Steele (2001a, 2001b)

[10] They suggest using $\bar{k}/K$ as prior probability for each variable where $\bar{k}$ is the average model size and K the total number of possible regressors

In the generic case (3) the BACE test has the following properties in the population:

	Truth: $\beta_1 = 0$	*Truth:* $\beta_1 \neq 0$
Correct	$if \quad \#S_1 = 0$	$if \quad \#S_1 = 1$ *or* $\#S_1 = K$
Incorrect	$if \quad \#S_1 = 1$ *or* $\#S_1 = K$	$if \quad \#S_1 = 0$
*Indeterminate**	$if \quad 2 \leq \#S_1 < K$	$if \quad 2 \leq \#S_1 < K$

* *depends on model priors, see proof*

where $\#\{\}$ *is the cardinality of a set.*

In the special case the test correctly determines the importance of the regressor.

Proof. See appendix. ■

In the generic case, the BACE does not work correctly. The reason is that BACE selects the best fitting model(s) (in the sense of minimizing $E(V(y \mid x_m, x_k)))$ among all the short regressions, which in general provides a biased estimate of β_1. In the special case, the model with x_1 and x_2 does provide the best fit and in this model $\gamma_1^{[12]}$ is unbiased for β_1. Thus, the identification is obtained using Theorem 3.

3.4 General-to-specific

The basic general-to-specific procedure has been refined by Hendry and Krolzig (2004) and Hoover and Perez (2004)[11]. The procedure starts with a general unrestricted model (called GUM) that cannot be rejected by a host of misspecification tests. Then the procedure searches over different paths, where models are restricted until all variables are significant. In the process the host of misspecification tests are applied which may also lead to backtracking on a path. In the end, a model is chosen that cannot be rejected by misspecification tests[12] and encompassing tests against other candidate models from other paths.

Because of the undersized sample a general unrestricted model cannot be estimated. Therefore, we perform general-to-specific on each short regression with the maximum number of regressors. Among the models selected by the general-to-specific procedure performed on these short regressions, the best one is chosen. The procedure[13] is shortly described below:[14]

[11] Hoover and Perez(1999) and Hendry and Krolzig(1999) also discuss refined versions of general-to-specific, but in a time-series context.

[12] Denoted a congruent model, see eg Hendry (1995)

[13] The procedure is similar to a procedure suggested by Hansen(1999) in a time-series context.

[14] There is no reference to misspecification tests since none of the short regressions here are misspecified.

a. Select a subset of K_s regressors, where K_s is the maximum number of regressors included in any short regression.

b. Perform general-to-specific. Delete the variable with the lowest insignificant t-statistics. Continue until all coefficients are significant.

c. Repeat a and b for all short regressions with K_s regressors.

d. Among the candidate models, choose one according to e.g. a model selection criteria. Here, choose the model with the lowest standard error. In case of a tie, x_1 is denoted important if it is included in one of the models in the tie.

The following proposition shows that the general to specific procedure determines importance of the variable of interest based on the best fitting model:

Proposition 7 (General-to-specific) *Let* $\sigma_{\min} = \min\limits_{i,j}(\sigma_{[ij]})$. *In the generic case (3) the extended general-to-specific procedure has the following properties in the population:*

	Truth: $\beta_1 = 0$	*Truth:* $\beta_1 \neq 0$
Correct	$\sigma_{\min} < \min\limits_{k=2,..,K} \sigma_{[1k]}$	$\sigma_{\min} = \min\limits_{k=2,..,K} \sigma_{[1k]}$
Incorrect	$\sigma_{\min} = \min\limits_{k=2,..,K} \sigma_{[1k]}$	$\sigma_{\min} < \min\limits_{k=2,..,K} \sigma_{[1k]}$

where $\sigma^2_{[mk]} = E(V(y \mid x_m, x_k))$.

In the special case the test correctly determines the importance of the regressor.

Proof. See appendix. ■

The result is similar to that of BACE. The general-to-specific procedure works in the special case because it is based on the criterion which insures identification according to Theorem 3.

3.5 Minimum t-statistic over models test

The minimum t-statistics over models test denotes the variable of interest as important if the minimum t-statistics (in absolute value) taken over short regressions including x_1 is statistical significantly different from 0.[15] This is similar to the modified version of the extreme bounds suggested earlier in this paper and in the spirit of the Sala-i-Martin (1997a, 1997b) test. White (2000) and Hansen (2003) have under different conditions shown that the bootstrap can be applied to approximate the distribution of the minimum t-statistics. The following proposition provides the properties of the minimum t-statistics over models test.

[15] Note, $P(|t_i| \ > c, \forall i) = P\left(\underset{i}{Min}|t_i| \ > c\right)$

Proposition 8 (Minimum t-statistics over models test) *In the generic case (3) the minimum t-statistics over models test has the following properties in the population:*

	Truth: $\beta_1 = 0$	*Truth:* $\beta_1 \neq 0$
Correct	*none*	*all cases*
Incorrect	*all cases*	*none*

In the special case the test correctly determines the importance of the regressor.
Proof. *See appendix.* ■

In the generic case, the minimum t-statistics over models test will not be consistent because it denotes the variable important in all cases. The reason why this happens when $\beta_1 = 0$ is that the coefficient on x_1 is different from 0 in all the short regressions because of the omitted variable bias. In the special case the regression of y on x_1 and x_2 is the correct regression and, thus, provides the correct information with respect to x_1. The other short regressions all have the coefficient to x_1 different from 0. It only takes one short regression with acceptance of $\beta_1 = 0$ to accept overall $\beta_1 = 0$ but all short regressions with rejection of $\beta_1 = 0$ to reject overall $\beta_1 = 0$. Thus, here the short regression of y on x_1 and x_2 in effect determines the outcome.

3.6 Model selection criteria: BIC and AIC

Model selection criteria are designed to select a model based on some criterion which is usually a penalized likelihood[16]. Then the significance of a particular variable can be assessed in the chosen model.

One model selection criterion is BIC (Schwarz information criterion). It turns out that the posterior probability of a model in the Bayesian averaging approach by Sala-i-Martin, Doppelhoffer and Miller (2004) can be rewritten as a function of BIC assuming conditional normality of y. BIC is given by:

$$BIC_j = N \log \frac{1}{N} SSE_j + \log (N) k_j,$$

where σ_j^2 is the maximum likelihood estimate of the variance of the error associated with model j and k_j is the number of parameters in model j.

The results in the next proposition show that BIC is similar to BACE and general-to-specific since these methods (in the population) select a model based on the same measure of fit.

[16] For an extended discussion see Burnham and Anderson(2002).

Proposition 9 (BIC) *In the generic case* (3) *BIC has the following properties in the population*:

	Truth: $\beta_1 = 0$	*Truth:* $\beta_1 \neq 0$
Correct	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} < \min_{k=2,..,K} \sigma_{[1k]}$	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} > \min_{k=2,..,K} \sigma_{[1k]}$
Incorrect	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} > \min_{k=2,..,K} \sigma_{[1k]}$	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} < \min_{k=2,..,K} \sigma_{[1k]}$
Indeterminate	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} = \min_{k=2,..,K} \sigma_{[1k]}$	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} = \min_{k=2,..,K} \sigma_{[1k]}$

where $\sigma^2_{[mk]} = E(V(y \mid x_m, x_k))$.

In the special case the test correctly determines the importance of the regressor.

Proof. See appendix. ■

The special case confirms that BIC is a consistent model selection criterion if the true model is included in the set of models investigated.

Another model selection criterion is the Akaike information criterion, AIC, and its corrected version, AICC. The AIC and AICC for a model j are given by:

$$\begin{aligned} AIC_j &= N \log \frac{1}{N} SSE_j + 2k_j \\ AICC_j &= AIC_j + \frac{2k_j\,(k_j+1)}{N-k_j-1}. \end{aligned}$$

The next proposition shows that AIC and AICC have properties similar to BIC

Proposition 10 (AIC and AICC) *In the generic case* (3) *AIC and AICC have the following properties in the population*:

	Truth: $\beta_1 = 0$	*Truth:* $\beta_1 \neq 0$
Correct	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} < \min_{k=2,..,K} \sigma_{[1k]}$	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} > \min_{k=2,..,K} \sigma_{[1k]}$
Incorrect	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} > \min_{k=2,..,K} \sigma_{[1k]}$	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} < \min_{k=2,..,K} \sigma_{[1k]}$
Indeterminate	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} = \min_{k=2,..,K} \sigma_{[1k]}$	$\min_{i\neq 1,j\neq 1} \sigma_{[ij]} = \min_{k=2,..,K} \sigma_{[1k]}$

where $\sigma^2_{[mk]} = E(V(y \mid x_m, x_k))$.

In the special case the test correctly determines the importance of the regressor.

Proof. See appendix. ■

For the population the correction to AIC does not matter. Though AIC and BIC both provide the correct answer in the special case, they do so in different ways. Under the truth $\beta_1 = 0$, BIC selects the regression of y on x_2 with probability 1, whereas AIC selects both the regression of y on x_2 and the regression of y on x_1 and x_2 with positive probability.[17] In both regressions, however, the coefficient on x_1 equals 0. The reason is that AIC in the population has a positive probability of selecting models that nest the true model.

4 New test

The main result in section 3 is that none of the methods are consistent in the generic case. This has to be the case in view of the impossibility theorem (Theorem 1). The conditions of the impossibility theorem may not be true under all circumstances with an undersized sample. In this section, we construct a test to determine the importance of the variable of interest that builds directly on checking the conditions of the impossibility theorem.

The test is also applicable when $N > K$. In that case, none of the omitted variables are linear combinations of the included variables and the impossibility theorem does not apply. The conditions in the impossibility theorem are, however, sufficient to identify β_1 which is well known from the omitted variable bias problem.[18]

The test involves two steps. The first step is finding a sufficient number of variables that are orthogonal to a conditioning set with x_1. The second step is to regress y on this set of conditioning variables. A set of orthogonal regressors can be found (if it exists) simply from the correlation matrix of the regressors. Such an approach avoids a curse of dimensionality.

The following algorithm is a practical way to implement the test. Let K_s be the number of variables allowed in a regression. ($K_s < N$).

1. Find all regressors correlated with x_1 and insert them in the set I^1. Stop if $\#I^1 > K_s$.[19] Let $m = 1$ and $j = 1$.

2. Let z_j be the j'th variable in I^m. Find the set of regressors, C, not in I^m that is correlated with z_j. Set $I^{m+1} = I^m \cup C$. Stop if $\#I^{m+1} > K_s$.

3. Set $m = m + 1$ and $j = j + 1$. Repeat 2 if $j \leq \#I^m$ otherwise continue with 4.

4. Regress y on the regressors in I^m and test if the coefficient on x_1 is different from 0.

[17] This confirms the known result that AIC is inconsistent if the true model is nested in some of the models investigated.

[18] In a well-posed linear regression problem with omitted variables, inclusion of the conditional expectations of the omitted variables insures identification of β_1 if the conditional expectations are different from a linear index.

[19] # is the cardinality (number of elements) in a set

To find uncorrelated regressors (in 1 and 2), a simple test for a zero correlation between two regressors can be used. For example, if joint normality between two regressors is assumed and ρ is the correlation coefficient, then $\sqrt{n-2}\rho/\sqrt{1-\rho^2}$ follows a t-distribution with $(n-2)$ degrees of freedom. In step 4 a t-test can be used.

The next theorem shows that the test defined by the steps 1 to 4 is consistent.

Theorem 11 (Test of importance) *The test, T, specified by the steps 1 to 4 is a consistent test of:*

I. Importance of x_1 can be determined

II. Under I, whether x_1 is important or not.

Proof. See appendix. ■

Property I tests if the problem is ill-posed or can be made well-posed. Property I can be tested by other approaches than the one outlined above the Theorem. For example, instead of investigating the correlation matrix between the regressors, the problem of finding orthogonal regressors can be formulated in a SURE system. After searching for orthogonal regressors, the set of conditioning regressors, I^m, may contain fewer members than than K_s. This implies that the regression in 4 can be extended with extra variables. In theory, it does not matter, but the finite sample properties may differ.

In the special case (4) in section 3, the new test should not be used. The test is not designed to this particular situation, where the true model is one of the short regressions, and many of the other existing methods are consistent in the special case as we have also proved in Section 3.

5 Finite sample properties of tests

In this section we investigate the finite sample properties of the new test and some of the methods considered in section 3. We focus on two setups which match the identification conditions in Theorem 1 and 3.

The finite sample properties are investigated by a Monte Carlo study. The design of the Monte Carlo study has four variables and a constant. The number of MC replications is 10,000. All variables have zero mean and unit variance. The purpose is to determine if the regressor x_1 is important. The design has x_1 and x_2 correlated with a correlation coefficient equal to 0.50, and they are uncorrelated with x_3 and x_4 which are also mutually uncorrelated. They are drawn from a multivariate normal distribution. The constant is equal to 2 and the random disturbance terms are iid $N(0, 4)$.

In the tables 1,2 and 3 below, the first four methods use all subsets of regressors with two variables (and a constant) at the most. The first four methods are calculated as described in section 3 with the exception that the model selection with General to specific is done with BIC. For the Bayesian tests, we apply a "significance" test on the unconditional posterior mean.

Tables 1 and 2 show for $n = 25$ and $n = 50$, respectively, a case where the true model is not included among the short regressions but x_3 and x_4 are orthogonal on x_1 and x_2. Hence, the conditions in the impossibility theorem fail. The coefficient vector on x_2, x_3 and x_4 is $(5, 5, -5)'$. The first four tests have a Type I error (columns with $\beta_1 = 0$) close to 0, because all models with x_1 fit worse than combinations of two of the three other variables. The test of importance has a Type I error close to 5%. Sala-i-Martin's test includes x_1 in all short regressions, and therefore some of the models produce biased estimates. Sala-i-Martin's test has a large probability of a Type I error because a high weight is put on models with good fit. In those models the estimate of x_1 is biased. The EBA has a low rejection error because the EBA only rejects the null when all variables have the same sign[20].

Many of the tests have low power. For a fairly wide interval of values of β_1, they have zero power. This is in accordance with the propositions in section 3. The test of importance has good power properties.

Table 1. Probability of denoting x_1 important
Design: $n = 25$ and $(\beta_2\beta_3\beta_4) = (5, 5, -5)$

	True values of β_1					
Methods	-10	-8	-6	-4	-2	0
AICC	0.9698	0.7905	0.3352	0.0396	0.003	0.0069
BIC	0.9698	0.7905	0.3352	0.0396	0.003	0.0069
GSP	0.9698	0.7905	0.3352	0.0396	0.003	0.0069
Bayes (uncond. posterior est.)	0.9122	0.6133	0.1668	0.0104	0.0001	0.0011
Sala-i-Martin's test	0.9959	0.9663	0.7563	0.2868	0.0941	0.3051
EBA	0.9867	0.8844	0.4962	0.0748	0.002	0.0229
Test of importance	0.9963	0.9748	0.8679	0.5514	0.1589	0.0401

Note: All significance tests use a nominal level of 5%.

[20] The EBA is carried out on the bounds as done in the literature. This amounts to using a critical value of 2 for the tests.

Table 2.Probability of denoting x_1 important

Design: $n = 50$ and $(\beta_2\beta_3\beta_4) = (5, 5, -5)$

	True values of β_1					
Methods	-10	-8	-6	-4	-2	0
AICC	0.9939	0.8274	0.194	0.0026	0.0000	0.0003
BIC	0.9939	0.8274	0.194	0.0026	0.0000	0.0003
GSP	0.9939	0.8274	0.194	0.0026	0.0000	0.0003
Bayes (uncond. posterior est.)	0.9886	0.7535	0.1229	0.001	0.0000	0.0000
Sala-i-Martin's test	1	0.9993	0.9483	0.4163	0.119	0.5517
EBA	1	0.997	0.8643	0.1685	0.001	0.0240
Test of importance	1	0.9998	0.9945	0.8803	0.3544	0.0456

Note: All significance tests use a nominal level of 5%.

Several of the tests have non-monotonic power. For example, Sala-i-Martin's test has the lowest power for $\beta_1 = -2$. One reason for this is that for $\beta_1 = -2$ the omitted variable bias in the estimator of β_1 in the best fitting model is about 0.

For some of the methods, the power decreases in some intervals of β_1 as the sample size increases. For example for $\beta_1 = -6$, the general-to-specific method has power 0.3352 for $n = 25$ and 0.194 for $n = 50$. The same happens for the BACE methods. The reason is that the best fitting model for $\beta_1 = -6$ does not include x_1, and when n increases this model is selected with higher probability.

Despite the fact that the conditions in the impossibility theorem are not satisfied for the design in tables 1 and 2 and, thus, the problem is well-defined, the existing methods (considered in section 3) do not work properly. For most of the methods, their selection are based on fit. Therefore, if the model with x_1 and x_3 provides the best fit, then the estimate of the coefficient will be biased due to the omitted variable bias caused by x_2.

The existing methods do not work correctly for an important regressor even if all the regressors are orthogonal. Again, the cause is that the selection of model is based on fit. Even though the estimate of the coefficient of x_1 is not biased, the methods may choose models where x_1 is not included. This suggests a modification to the methods running all short regressions. They can be made consistent when all the regressors are orthogonal by only considering short regressions, which include x_1.

Table 3 shows the result of a design similar to the special case in section 2 ($\beta_3 = \beta_4 = 0$). Thus, the long regression is included as one of the short regressions. As expected, the methods perform well. Most of the methods from section 2 have better power than the test of importance. Unfortunately, it is not possible to determine if the special case is true, that is, if $\beta_3 = \beta_4 = 0$. Therefore, it may not be unreasonable to suspect that the test of importance performs less well in the special case since it is consistent in a wider class of models.

Table 3. Probability of denoting x_1 important
Design: $n = 25$ and $(\beta_2\beta_3\beta_4) = (5, 0, 0)$.

	True values of β_1						
Methods	-6	-4	-2	-1.5	-1	-0.5	0
AICC	1	1	0.9999	0.952	0.8126	0.1651	0.0473
BIC	1	1	0.9999	0.952	0.8126	0.1651	0.0473
GSP	1	1	0.9999	0.952	0.8126	0.1651	0.0473
Bayes (uncond. posterior est.)	1	1	0.9992	0.7595	0.458	0.0266	0.0027
Sala-i-Martin's test	1	1	0.9999	0.9656	0.8491	0.1972	0.0626
EBA	0.9953	0.884	0.2936	0.006	0.0015	0.001	0.0251
Test of importance	0.9995	0.9961	0.9638	0.8686	0.7448	0.158	0.0548

Note: All significance tests use a nominal level of 5%.

6 Effectiveness of foreign aid

Many papers investigate the important question of the effectiveness of foreign aid on growth. One question is if "good policies" by the receiving country significantly influences the effect of foreign aid. A leading example is the paper by Burnside and Dollar (2000). They find that good policy has a significant impact on the effectiveness of foreign aid. Their findings, however, have be questioned by, for instance, Easterly, Levine and Roodman (2004) for data reasons. We investigate the question in view of the fact that many other variables are found to be important for growth.

The set of variables believed to be of importance is determined as follows. Burnside and Dollar (2000) have panel data for 56 countries from 1970-1993. Growth rates are calculated as averages over four years. Their most extensive specification includes the variables from the source Burnside and Dollar (2000) listed in table 4. They also include time dummies. A number of other variables have been suggested in the literature (see Roodman (2004) for a summary of the literature). These variables have been collected by Roodman (2004) and the most important ones are listed in the table. Among the variables included in the table are those used by Dalgaard, Hansen and Tarp (2004), who suggest an alternative aid interaction model based on tropical area.

We use the new test to investigate the importance of good policy for aid effectiveness. This is modelled as the interaction between the good policy index[21] and aid. In total 10

[21] This index is calculated as:
$1.28 + 6.85 budget\ surplus - 1.4 * inflation + 2.16 * opennes$

variables and time and country dummies are included. Thus, to falsify the conditions of the impossibility theorem, 8 variables must be orthogonal on the remaining 10 variables of which one is the interaction between aid and the good policy index.

Table 4: Variables used in aid-growth models

Variable	Source
Growth rate of GDP per capita, 4 year averages	Burnside & Dollar (2000)
Good Policy index	Burnside & Dollar (2000)
Effective development aid as percentage of GDP	Burnside & Dollar (2000)
Assassinations per capita	Burnside & Dollar (2000)
Ethnic fractionalization	Burnside & Dollar (2000)
Assassinations per capita*Ethnic fractionalization	Burnside & Dollar (2000)
Institutional quality	Burnside & Dollar (2000)
M2 as percentage of GDP (lagged)	Burnside & Dollar (2000)
Initial log GDP per capita	Burnside & Dollar (2000)
Aid*Good policy	Burnside & Dollar (2000)
East Asian dummy	Burnside & Dollar (2000)
Africa dummy	Burnside & Dollar (2000)
Aid squared*policy	Burnside & Dollar (2000)
Mean years of secondary schooling among those over 25	Roodman (2004)
(Log) Population	Roodman (2004)
Population growth	Roodman (2004)
Political instability (lagged)	Roodman (2004)
Tropical area (as a fraction of total area)	Roodman (2004)
Positive Shock to export prices	Roodman (2004)
Negative Shock to export prices	Roodman (2004)
Aid*tropical area	Roodman (2004)

The new test involves multiple testing. Both relative conservative and liberal critical values are investigated. The conservative critical value is selected by ignoring the multiple testing problem, and using a 5% nominal level for each test of zero correlation. With this critical value there are not sufficiently many regressors which were orthogonal.[22] The liberal critical value is based on the Bonferroni bound[23]. Assuming that there are 8 orthogonal regressors, the null is rejected if the p-value associated with the test is less

[22] Using the regular 5% cut-off corresponding to a critical value of 1.97, ten regressors are left after the first round and as some of the regressors remaining are correlated with aid*tropics, the search ends.

[23] Note, usually the Bonferroni bound is conservative but here the Bonferroni bound is used in a two-step procedure; The smaller critical value, the less likely orthogonal regressors can be found.

than 0.05/80 corresponding to a critical value of about 3.47 in absolute value[24]. The result is that based on the Bonferroni bound based critical value, there are not enough orthogonal regressors.[25]

In conclusion, with either choice of critical values there is clear evidence that not enough orthogonal regressors exist. Therefore, if many regressors are believed to be of importance, then the empirical model in Burnside and Dollar (2000) cannot be used to determine if aid is more effective when combined with good policy.

7 Conclusion

We considered the economic task of determining if a variable is important in a regression with more variables believed to be of importance than observations. We found conditions under which the undersized sample leads to an ill-posed inverse problem. The problem can only be well-posed if there is a sufficient number of orthogonal regressors.

In light of the impossibility of the task, it came as no surprise that existing model selection methods cannot resolve the ill-posed inverse problem. The analysis showed that the majority of these methods are based on a measure of model fit. Therefore, these methods do not work even when the problem is well-posed. The task can also be interpreted as inferring the effect of a variable when the true model is (infinitely) larger than models that can be estimated.

We derived consistency properties of commonly applied methods by explicitly taking the ill-posed problem caused by an undersized sampled into account. The results are very different from standard asymptotic results developed for the methods. In fact, when the problem is ill-posed none of the methods are consistent. The analysis of the ill-posed problem also provides insight into the properties of the methods when they are applied to well-posed problems. This can be useful when the degrees of freedom is low.

Fundamentally, our results illustrate the importance of choosing a loss function appropriate for the task at hand. The model selection methods build on loss functions that are based on measures of model fit. These are appropriate when the true model can be estimated. In the general case which is considered here, a loss function based on model fit is problematic as proved by the impossibility theorem. A loss function based on failing the conditions of the impossibility theorem seems more promising.

[24] Under the null there are 80 possible tests in which rejection could happen erroneously. Using a critical value of 0.05/80 puts a bound on the type I error. Given that the test is two-sided, we choose 3.47 as the critical value.

[25] In the first round, four regressors were found to be correlated with aid*good policy one of which is aid*tropical area. This variable is correlated with an additional five regressors one of which is log GDP. Finally in the third round one regressor is correlated with log GDP and the search stops.

The identification results are used on the problem of determining if aid is more effective when combined with good policy. We showed that this question cannot be answered by using a model of the type suggested by Burnside and Dollar (2000) in the presence of many variables believed to be important for growth without adding further information.

8 Appendix

Proof of Theorem 1 (Impossibility of determining importance). The defining property of the undersized sample problem is a reduced rank of the regressor matrix. Here, the rank is N $(< K)$. The implication of a reduced rank is that N of the regressors can span a space that includes the remaining $(K - N)$ regressors.

With probability 1, any subset of N regressors can span the space. Consider regressor 1 to N. In the undersized sample,

$$\Pr\left(rank \begin{bmatrix} X_{11} & .. & X_{N1} \\ : & & : \\ X_{1N} & .. & X_{NN} \end{bmatrix} = N \right) = 1, \tag{6}$$

since it is assumed that the regressors in the population are not linear dependent. Let $\underline{X}_k = (X_{k1}, .., X_{kN})'$ be the values of the k'th regressor. Using the first N regressors to span the space, the remaining regressors can be written as:

$$\underline{X}_i = \sum_{k=1}^{N} a_k^i \underline{X}_k, \ i = N+1, .., K,$$

where a_k^i are random variables determined by the following system:

$$a^i \equiv \begin{bmatrix} a_1^i \\ : \\ a_N^i \end{bmatrix} = \begin{bmatrix} X_{11} & .. & X_{N1} \\ : & & : \\ X_{1N} & .. & X_{NN} \end{bmatrix}^{-1} \begin{bmatrix} X_{i1} \\ : \\ X_{iN} \end{bmatrix} \quad i = N+1, .., K. \tag{7}$$

Because of the reduced rank,

$$E(\underline{y} \mid \underline{X}_1, .., \underline{X}_N, \underline{X}_{N+1} .., \underline{X}_K) = E(\underline{y} \mid \underline{X}_1, .., \underline{X}_N),$$

where

$$E(\underline{y} \mid \underline{X}_1, .., \underline{X}_N) = \sum_{k=1}^{N} \beta_k \underline{X}_k + \sum_{i=N+1}^{K} E\left((\beta_i \sum_{k=1}^{N} a_k^i \underline{X}_k) \mid \underline{X}_1, .., \underline{X}_N \right).$$

This can be rewritten as

$$\begin{aligned} E(\underline{y} \quad \mid \quad \underline{X}_1, .., \underline{X}_N) &= \sum_{k=1}^{N} \left(\beta_k + \sum_{i=N+1}^{K} \beta_i E\left(a_k^i \mid \underline{X}_1, .., \underline{X}_N\right) \right) \underline{X}_k \qquad (8) \\ &= \sum_{k=1}^{N} \left(\beta_k + \sum_{i=N+1}^{K} \beta_i \mu_k^i \right) \underline{X}_k, \end{aligned}$$

where $\mu_k^i = E\left(a_k^i \mid \underline{X}_1, .., \underline{X}_N\right)$. It can be seen that the coefficient on each $\underline{X}_k$ is a linear combination of the coefficient, β_k, from the long regression in the population and terms from the constraint leading to the reduced rank. The system (8) has N equations with K unknown parameters.

Recovery of β_1 in the system (8) of conditional expectations is possible if

$$E(\underline{y} \mid \underline{X}_1, .., \underline{X}_N; \beta_1, .., \beta_K) \neq E(\underline{y} \mid \underline{X}_1, .., \underline{X}_N; \beta_1^*, .., \beta_K^*)$$

for any choice of $\beta_1^* \neq \beta_1$ and $\beta_2^*, .., \beta_K^*$. Let the coefficient to $\underline{X}_k$ be $c_k (= \beta_k + \sum\limits_{i=N+1}^{K} \beta_i \mu_k^i)$. Then the condition fails if an $\beta_1^* \neq \beta_1$ can be found so that $c_k = c_k^*$. In matrix form:

$$\begin{bmatrix} 1 & & & \mu_1^{N+1} & \cdots & \mu_1^K \\ & \ddots & & \vdots & \ddots & \vdots \\ & & 1 & \mu_K^{N+1} & \cdots & \mu_K^K \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \\ \beta_{N+1} \\ \vdots \\ \beta_K \end{bmatrix} = \underline{c}, \tag{9}$$

where $\underline{c} = (c_1, .., c_N)'$. The degrees of freedom to determine β_1 only depends on $\mu_1^{N+1}, .., \mu_1^K$ since there are no restrictions imposed on $\beta_{N+1}, .., \beta_K$ from the equations involving $\beta_2, .., \beta_N$. Thus, only if all $\mu_1^{N+1}, .., \mu_1^K$ are equal to 0, then $\beta_1^* = \beta_1$ is a unique solution.

The values of $\mu_1^{N+1}, .., \mu_1^K$ are determined from the regressions of the omitted variables on the included variables. Using (7) and denoting by $\underline{d}_k$ the k'th row of the inverse matrix, get:

$$\mu_k^i = E\left(a_k^i \mid \underline{X}_1, .., \underline{X}_N\right) = \underline{d}_k \cdot E\left(\underline{X}_i \mid \underline{X}_1, .., \underline{X}_N\right).$$

Using (6), with probability 1, μ_k^i equals 0 if and only if $E\left(\underline{X}_i \mid \underline{X}_1, .., \underline{X}_N\right) = 0$. Thus, without imposing restrictions on the β's, β_1 can only be recovered if $E\left(\underline{X}_i \mid \underline{X}_1, .., \underline{X}_N\right) = 0$ for all $i = N+1, .., K$.

The argument above is for the set of N included variables $A = \{X_2, .., X_N\}$ together with X_1 and the set of $(K - N)$ excluded variables $B = \{X_{N+1}, .., X_K\}$. The argument can be repeated for any combinations of included and excluded variables. Therefore, for a recovery of β_1 there needs to exist at least one set A (and B) so that the corresponding μ_1's all equal 0. ■

Proof: Theorem 2 (Partial identification of variables of importance). The result (9) in the proof of theorem 1 can be used to show partial identification of importance for a set of regressors. Consider identification of at least one of the regressors in the set

$R = \{X_1, .., X_r\}$ as important (not which one of them). Suppose none of these regressors are important, $\beta_1 = .. = \beta_r = 0$. Then from (9) get

$$\begin{bmatrix} 1 & & & \mu_1^{N+1} & \cdots & \mu_1^K \\ & \ddots & & \vdots & \ddots & \vdots \\ & & 1 & \mu_K^{N+1} & \cdots & \mu_K^K \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta_{r+1} \\ \vdots \\ \beta_{N+1} \\ \vdots \\ \beta_K \end{bmatrix} = \underline{c},$$

The first r rows of this linear system can be solve for $\beta_{N+1}, .., \beta_K$ if $(K - N) \leq r$. If $(K - N) < r$, then there are more equations (restrictions) than parameters. Only if $\beta_1 = .. = \beta_r = 0$ is correct do these equations have a solution. This insures the identification.

The number of excess variables $(K - N)$ may be greater than r if enough of them are orthogonal to the variables in the importance set R. For example, if $\mu_1^j = .. = \mu_r^j = 0$ for some $j \in \{N + 1, .., K\}$, then β_j can be ignored. From the proof of theorem 1, $\mu_i^j = 0$ if $E\left(\underline{X}_i \mid \underline{X}_1, .., \underline{X}_N\right) = 0$. Therefore, for identification less than r of the excess variables must be correlated with the variables in R. ■

Proof: Theorem 3 (Identification assuming a true submodel). The result is a generalization of e.g. Wooldridge (2002), p. 31, property CV.3. For any subset $C \subset \{x_1, .., x_K\}$,

$$E_{x_1,..,x_K}\left(V(y \mid x_1, .., x_K)\right) = E_C\left(V(y \mid C)\right) - E_{x_1,..,x_K}\left(E(y \mid x_1, .., x_K) - E_C(y \mid C)\right)^2.$$

It follows that

$$\begin{array}{ll} 1) & \text{If } E(y \mid x_1, .., x_K) = E_C(y \mid C)\text{, then } E\left(V(y \mid x_1, .., x_K)\right) = E\left(V(y \mid C)\right) \\ 2) & \text{If } E(y \mid x_1, .., x_K) \neq E_C(y \mid C)\text{, then } E\left(V(y \mid x_1, .., x_K)\right) < E\left(V(y \mid C)\right). \end{array} \qquad (10)$$

Therefore, if A is the smallest subset such that $E(y \mid x_1, .., x_K) = E_A(y \mid A)$, then 2) holds for any set B which does not contain A as a subset. ■

Proof of Proposition 4 (Extreme Bounds). In the sample, x_1 is robust if the estimates of the coefficient to x_1 in all the short regressions are significant and have the same sign. In the population with no estimation uncertainty, the conditions for characterizing a variable as not important if and only if a coefficient to x_1 is 0 ($\gamma_1^{[1k]} = 0$ for some k) or there is a sign change ($Sgn(\gamma_1^{[1k]}) \neq Sgn(\gamma_1^{[1m]})$ for some k, m).[26]

In the generic case, (3), $\gamma_1^{[1k]} \neq 0$ for all k. Hence, x_1 is denoted not important if and only if there is a change of sign of the coefficients over the short regressions. Under the truth $\beta_1 = 0$, this is determined by the sign of the terms $c'_{[12]}\beta_{3;4}$, $c'_{[13]}\beta_{2;4}$ and $c'_{[14]}\beta_{2;3}$. Under the truth $\beta_1 \neq 0$, the true value of β_1 is important. By noting that the last condition for characterizing the variable as not important can be rewritten as $\underset{k}{Min}\left(\gamma_1^{[1k]}\right) < 0 < \underset{k}{Max}\left(\gamma_1^{[1k]}\right)$, the result of the proposition is achieved by making a substitution for the γ's.

In the special case, (4), $\gamma_1^{[12]} = \beta_1$. Hence, under the truth $\beta_1 = 0$, the condition for not important is satisfied. Under the truth $\beta_1 \neq 0$, none of the γ_1's equal 0. ■

Proof of Proposition 5 (Sala-i-Martin's test). In proving this and the following propositions, the next property of conditional variances is useful

$$\begin{array}{ll} 1) & \text{If } E(y \mid x_1, x_2) = E(y \mid x_2), \text{ then } E\left(V(y \mid x_1, x_2)\right) = E\left(V(y \mid x_2)\right) \\ 2) & \text{If } E(y \mid x_1, x_2) \neq E(y \mid x_2), \text{ then } E\left(V(y \mid x_1, x_2)\right) < E\left(V(y \mid x_2)\right). \end{array} \tag{11}$$

This is a special case of the result shown in the proof of Theorem 2.

The decision is based on whether $CDF(0) = \sum_{i=1}^{m} w_i CDF_i(0)$ is above or below $1 - \alpha$, where α resemblances a significance level. The Sala-i-Martin's test does not have an obvious analogue in the population, and therefore the population version is derived next as a probability limit.

Firstly, consider $CDF_{[1k]}(0) = Max\left(\Phi(\widehat{\gamma}_1^{[1k]}/\widehat{\sigma}_{\widehat{\gamma}_1^{[1k]}}), 1 - \Phi(\widehat{\gamma}_1^{[1k]}/\widehat{\sigma}_{\widehat{\gamma}_1^{[1k]}})\right)$. In the population γ is known and there is no uncertainty. If $\gamma_1^{[1k]} \neq 0$, then $CDF_{[1k]}(0) = 1$. If $\gamma_1^{[1k]} = 0$, then both the numerator and the denominator equal 0. Under suitable regularity conditions $\widehat{\gamma}_1^{[1k]}/\widehat{\sigma}_{\widehat{\gamma}_1^{[1k]}} \to^p Z$, $Z \sim N(0,1)$. Since $\Phi(Z) \sim U$, $U \sim Uniform[0,1]$,

$$P(CDF_{[1k]}(0) < a \mid \gamma_1^{[1k]} = 0) = P(Max(U, 1-U) < a) = 2a - 1, \; 0.5 \leq a \leq 1. \tag{12}$$

Secondly, the weight can be rewritten as

$$w_j = \frac{SSE_j^{-N/2}}{\sum_{i=1}^{m} SSE_i^{-N/2}} = \frac{1}{\sum_{i=1}^{m} \left(\frac{\frac{1}{N}SSE_i}{\frac{1}{N}SSE_j}\right)^{-\frac{N}{2}}}, \tag{13}$$

[26] In terms of the t-statistics and asymptotics, the decision rule can be determined the following way. The t-statistics used for testing $\gamma_1^{[1k]} = 0$ is given by $\widehat{t}_1^{[1k]} = \widehat{\gamma}_1^{[1k]}/\sqrt{V(\widehat{\gamma}_1^{[1k]})}$, where ^indicates the estimator, for instance the OLS estimator. The probability limit of the t-statistics is degenerate at +∞ or -∞ when $\gamma_1^{[1k]}$ is positive or negative, respectively (consistency of t-test). For $\gamma_1^{[1k]} = 0$ the distribution of the t-statistics is $N(0,1)$ under regularity conditions. When the sample size approach ∞, the significance probability should approach 0 and, thus, the probability of accepting approaches 1.

where SSE_j is the sum of squared residuals in regression j. In a regression of y on x_1 and x_k

$$\frac{1}{N}SSE_{[1k]} \rightarrow^p E\left((y-(x_1\gamma_1^{[1k]}+x_k\gamma_k^{[1k]}))^2\right) = E_{x_1,x_k}(V(y \mid x_1,x_k)) \equiv \sigma^2_{[1k]}$$

under suitable regularity conditions. The first equality sign comes from the assumption that all short regressions are linear.

The convergence of the terms $\left(\frac{1}{N}SSE_{[1k]}/\frac{1}{N}SSE_{[1i]}\right)^{\frac{N}{2}}$ depends on the probability limits of the numerator and denominator. When they are different

$$\left(\frac{\frac{1}{N}SSE_{[1k]}}{\frac{1}{N}SSE_{[1i]}}\right)^{\frac{N}{2}} \rightarrow^p \begin{cases} \infty & \text{if } \sigma_{[1k]} > \sigma_{[1i]} \\ 0 & \text{if } \sigma_{[1k]} < \sigma_{[1i]} \end{cases}.$$

If the probability limits of the numerator and denominator are equal, the ratio raised to the power of N may converge to a random variable with a non-degenerate distribution. In order to see this, note that for regular problems where the coefficient estimators are $\sqrt{N}$ consistent,

$$N\left(\log(\frac{1}{N}SSE_{[1k]}) - \log(\sigma^2_{[1k]})\right) \rightarrow^d W_{[1k]},$$

where $W_{[1k]}$ has a non-degenerate distribution. A similar result holds for the regression of y on x_1 and x_i

$$N\left(\log(\frac{1}{N}SSE_{[1i]}) - \log(\sigma^2_{[1i]})\right) \rightarrow^d W_{[1i]}.$$

By subtracting these two and taking the exponential the result is:

$$\left(\frac{\frac{1}{N}SSE_{[1k]}}{\frac{1}{N}SSE_{[1i]}}\right)^{\frac{N}{2}} \rightarrow^d \left(e^{(W_{[1k]}-W_{[1i]})}\right)^{1/2} = \left(e^{W_{[1k]/[1i]}}\right)^{1/2}. \tag{14}$$

This is a non-degenerate distribution. Note that the case with one regression nested in the other say $k=0$, $\log\left(\frac{1}{N}SSE_{[10]}/\frac{1}{N}SSE_{[1i]}\right)^N$ is the LR test statistic. Under the null hypothesis (equivalent to $\sigma_{[1i]} = \sigma_{[1i]}$), $W_{[1k]/[1i]}$ is a chi-square with 1 degrees of freedom distributed random variable.

In the generic example, the weight in the population can be written as:

$$w_{[1k]} = \underset{N\rightarrow\infty}{plim} \frac{1}{1+\sum\limits_{i=2,i\neq k}^{m}\left(\frac{\frac{1}{N}SSE_{[1k]}}{\frac{1}{N}SSE_{[1i]}}\right)^{\frac{N}{2}}}.$$

The limit of the summation in the denominator is determined by the size of $\sigma^2_{[1k]}$ relative to $\sigma^2_{[1i]}$. Let $\#\{\}$ be the cardinality of a set. Then $c = \#\{k \mid \sigma_{[1k]} = \min\limits_i(\sigma_{[1i]})\}$ is the number of variances achieving the lowest value. The weight in the population can be expressed the following way:

$$w_{[1k]} = \begin{cases} 0 & \text{if } \sigma_{[1k]} > \min\limits_{i\neq k}(\sigma_{[1i]}) \\ 1 & \text{if } \sigma_{[1k]} < \min\limits_{i\neq k}(\sigma_{[1i]}) \\ W^* & \text{if } \sigma_{[1k]} = \min\limits_{i\neq k}(\sigma_{[1i]}) \end{cases}, \tag{15}$$

where W^* is a random variable determined as a function of random variables from (14). Therefore, is has a non-degenerate distribution with support on a subset of $(0,1)$.

The value of $CDF(0)$ can now be determined in the generic case, (3). In the generic case, $\gamma_1^{1k} \neq 0$ for all k and, thus, $CDF(0) = \sum_{i=1}^{m} w_{[1i]} 1 = 1$. Therefore, the variable x_1 is denoted important no matter if $\beta_1 = 0$ or $\beta_1 \neq 0$.

In the specific case, (4), $\gamma_1^{[1k]} \neq 0$ for all k if $\beta_1 \neq 0$. The conclusion follows as above and x_1 is denoted important. If $\beta_1 = 0$ in the specific case, then $\gamma_1^{[12]} = 0$ and $CDF(0) = w_{[12]}CDF_{[12]}(0) + \sum_{i=3}^{4} w_{[1i]}CDF_{[1i]}(0)$. According to (15) the conditional variances determine the weights. In this case $\sigma_{[12]} < \sigma_{[1k]},\ k > 2$ according to 2) in (10) since $E(V(y \mid x_1, x_2, x_3, x_4)) < E(V(y \mid x_1, x_k))$ and $E(V(y \mid x_1, x_2, x_3, x_4)) = E(V(y \mid x_1, x_2)) = \sigma_{[12]}$ $(= E(V(y \mid x_2)))$. Therefore, $w_{[12]} = 1$. In the population, the significance level can be set at $\alpha = 0$. Thus, using (12) the probability of denoting x_1 as not important is $P(CDF(0) < 1 \mid \gamma_1^{[12]} = 0) = 1$. ■

Proof of Proposition 6 (BACE). The decision is based on the expected value of γ_1 taken over models. The model posterior, (5), can be rewritten as

$$P(M_j \mid y) = \frac{1}{1 + \sum_{i=1, i \neq j}^{C^*} \frac{P(M_i)}{P(M_j)} N^{(k_j - k_i)/2} \left(\frac{1}{N}SSE_i / \frac{1}{N}SSE_j\right)^{-\frac{N}{2}}}.$$

The population analog can be derived as the probability limit for $N \to \infty$. In the regression of y on x_m and x_k, $\frac{1}{N}SSE_{[mk]} \to^p E_{x_m, x_k}(V(y \mid x_m, x_k)) \equiv \sigma^2_{[mk]}$, see the proof of proposition 4. The BACE method includes all short regressions. For notational convenience, let 0 in the index denotes no variable, for example, [30] is a regression of y on x_3. The model posterior in the population can be written as:

$$P(M_{[mk]} \mid y) = \underset{N \to \infty}{plim} \frac{1}{1 + \sum_{i=0}^{4} \sum_{j=i+1}^{4} \frac{P(M_{[ij]})}{P(M_{[mk]})} N^{\left(k_{[mk]} - k_{[ij]}\right)/2} \left(\frac{\frac{1}{N}SSE_{[ij]}}{\frac{1}{N}SSE_{[mk]}}\right)^{-\frac{N}{2}}}.$$

The limit properties of $\left(\frac{1}{N}SSE_{[ij]} / \frac{1}{N}SSE_{[mk]}\right)^{-\frac{N}{2}}$ can be determined in a similar manner as done in the proof of proposition 5. Therefore,

$$\left(\frac{\frac{1}{N}SSE_{[mk]}}{\frac{1}{N}SSE_{[ij]}}\right)^{\frac{N}{2}} \to^d \begin{cases} 0 & \text{if } \sigma_{[mk]} > \sigma_{[ij]} \\ \infty & \text{if } \sigma_{[mk]} < \sigma_{[ij]} \\ W_{[mk]/[ij]} & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \end{cases},$$

where $W_{[mk]/[ij]}$ is a random variable with a non-degenerate distribution and a support on a subset of $(0,\infty)$. This implies that

$$N^{\left(k_{[mk]}-k_{[ij]}\right)/2}\left(\frac{\sigma^2_{[mk]}}{\sigma^2_{[ij]}}\right)^{\frac{N}{2}} \underset{N\to\infty}{\to^d} \begin{cases} \infty & \text{if } \sigma_{[mk]} > \sigma_{[ij]} \\ 0 & \text{if } \sigma_{[mk]} < \sigma_{[ij]} \\ \infty & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} > k_{[ij]} \\ 0 & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} < k_{[ij]} \\ W_{[mk]/[ij]} & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} = k_{[ij]} \end{cases}.$$

Assuming that all models have a positive prior, the model posterior is for $m \neq 0$ and $k \neq 0$

$$P(M_{[mk]} \mid y) = \begin{cases} 0 & \text{if } \sigma_{[mk]} > \min\limits_{i,j\ (\neq m,k)}(\sigma_{[ij]}) \\ 1 & \text{if } \sigma_{[mk]} < \min\limits_{i,j\ (\neq m,k)}(\sigma_{[ij]}) \\ 0 & \text{if } \sigma_{[mk]} = \min\limits_{j}(\sigma_{[j0]}) \\ W_p & \text{if } \sigma_{[mk]} = \min\limits_{i,j\ (\neq m,k)}(\sigma_{[ij]}) > \min\limits_{j}(\sigma_{[j0]}) \end{cases},$$

where W_p is a random variable with a non-degenerate distribution and a support on a subset of $(0,1)$. For $k=0$

$$P(M_{[m0]} \mid y) = \begin{cases} 0 & \text{if } \sigma_{[m0]} > \min\limits_{i,j\ (\neq m,0)}(\sigma_{[ij]}) \\ 1 & \text{if } \sigma_{[m0]} = \min\limits_{i,j\ (\neq 0,0)}(\sigma_{[ij]}) > \min\limits_{j\neq m}(\sigma_{[j0]}) \\ W_{p1} & \text{if } \sigma_{[m0]} = \min\limits_{i,j\ (\neq 0,0)}(\sigma_{[ij]}) = \min\limits_{j\neq m}(\sigma_{[j0]}) \end{cases},$$

where W_{p1} is a random variable with a non-degenerate distribution and a support on a subset of $(0,1)$.

The value of γ_1 which is used for deciding the importance of x_1 is determined as the expected value over models:

$$\gamma_1^{SDM} = \sum_{i,j}\gamma_1^{[ij]}P(M_{[ij]} \mid y).$$

The decision rule is:

$$\begin{aligned} \text{Not important if } \gamma_1^{SDM} &= 0 \\ \text{Important if } \gamma_1^{SDM} &\neq 0 \end{aligned}$$

In the generic case, (3), it is only necessary to consider regressions with two regressors and the regression with x_1. Firstly to show that no regressions with a single regressor other than x_1 have a smaller variance, it is shown that $\sigma_{[1k]} < \sigma_{[k0]}$ for $k > 1$. Without loss of generality, assume that $E(y \mid x_3, x_4) = E(y \mid x_4)$. If $E(y \mid x_1, x_4) = E(y \mid x_1)$, then $E(y \mid x_1, x_4) \neq E(y \mid x_4)$ and, thus, $\sigma_{[14]} < \sigma_{[40]}$ using (11). The case of $E(y \mid x_1, x_4) = E(y \mid x_4)$, is ruled out by the assumptions of the generic case, (3). Finally, in the remaining cases $E(y \mid x_1, x_4) \neq E(y \mid x_4)$ and, thus, $\sigma_{[14]} < \sigma_{[40]}$ using (11). Secondly, $\sigma_{[10]}$ may have the smallest variance. In that case, $\sigma_{[10]} = \sigma_{[1k]}$ for all $k > 1$.

In the generic case it is only known that $E(V(y \mid x_1, x_2, x_3, x_4)) = E(V(y \mid x_2, x_3, x_4))$ when $\beta_1 = 0$. This is not useful in determining which model(s) are selected. There are four cases. If a short regression with x_1 has a unique smallest variance, say, $\sigma_{[1k]}$, then $\gamma_1^{SDM} = \gamma_1^{[1k]} \neq 0$. In the case that none of the short regressions with x_1 have the smallest variance, then γ_1^{SDM} is a linear combination of $\gamma_1^{[ij]} = 0$, $i, j > 2$, and thus $\gamma_1^{SDM} = 0$. If several short regressions with x_1 achieve the smallest variance and the regression with x_1 does not, a linear combination of the $\gamma_1^{[1k]}$'s with their priors may or may not result in $\gamma_1^{SDM} = 0$. Finally, if the regression with only x_1 achieves the smallest variance, then the posterior probability of that model is 1 since this is the only regression with one variable that can achieve the smallest variance. Note in this final case, $\sigma_{[10]} = \sigma_{[1k]}$, for all k.

In the special case and under the truth $\beta_1 = 0$, $E\left(V(y \mid x_1, x_2, x_3, x_4)\right) = E\left(V(y \mid x_2)\right)$. Thus, only short regressions including regressor x_2 will achieve the smallest variance. Since $\gamma_1^{[12]} = \gamma_1^{[2j]} = 0$ for any j, it can be concluded that $\gamma_1^{SDM} = 0$. Under the truth $\beta_1 \neq 0$, the short regression of y on x_1 and x_2 has the smallest variance, $\sigma_{[12]}$. For that regression, $\gamma_1^{[12]} \neq 0$ and, thus, $\gamma_1^{SDM} \neq 0$. ■

Proof of Proposition 7 (General-to-specific). The general-to-specific procedure on a short regression of y on x_m and x_k eliminates insignificant coefficients. In the population, the significance of the coefficients is determined by the values of $\gamma_m^{[mk]}$ and $\gamma_k^{[mk]}$. A regressor is only excluded if the corresponding coefficient is 0. Therefore, the variances $\sigma_{[mk]}$ can be used to determine the amount of reduction since $\sigma_{[mk]} = \sigma_{[m0]}$ if and only if $\gamma_k^{[mk]} = 0$.

In the generic case it was shown in the proof of proposition 4 that the only candidates for minimum variance are $\sigma_{[mk]}$, $m, k \geq 1$ and $\sigma_{[10]}$. Should there be a tie among several models, the regressor x_1 is denoted important if it was included in any of the models in the tie. Therefore, x_1 is denoted important if any $\sigma_{[1k]}$ achieves the lowest variance.

In the special case, the correctly specified regression is included in the short regressions. Hence, that regression will have the lowest variance. Under the truth $\beta_1 = 0$, the smallest variance is $\sigma_{[20]}$. This implies that $\sigma_{[2k]} = \sigma_{[20]}$ for all k and $\sigma_{[ij]} > \sigma_{[20]}$ for $i, j \neq 2$. The only regression with the smallest variance and x_1 is the regression of y on x_1 and x_2, but $\gamma_1^{[12]} = 0$. Under the truth $\beta_1 \neq 0$, the variance is uniquely minimized by $\sigma_{[12]}$. In this case, $\gamma_1^{[12]} \neq 0$. ■

Proof: Proposition 8 (Minimum t-statistics over models test).

The result follows from noting that the test will accept if any of the coefficients $\gamma_1^{[1k]}$ equals 0. In the generic case $\gamma_1^{[1k]} \neq 0$ for $k = 2, .., K$ and $\gamma_1^{[10]} \neq 0$. In the special case,

$\gamma_1^{[1k]} \neq 0$, $k = 3, .., K$ and $\gamma_1^{[10]} \neq 0$, but $\gamma_1^{[12]} = 0$ if $\beta_1 = 0$ and $\gamma_1^{[12]} \neq 0$ if $\beta_1 \neq 0$. Thus, the test provides the correct answer. ■

Proof: Proposition 9 (BIC). The choice of model can be determined by the differences in BIC. A model i is chosen over a model j if and only if

$$N(\log \frac{1}{N} SSE_i - \log \frac{1}{N} SSE_j) + \log(N)\,(k_i - k_j) < 0$$

for all $i, j \neq m, k$.

The population equivalent or probability limit of $\frac{1}{N} SSE_{[mk]}$ is $\sigma^2_{[mk]}$. In case $\sigma_{[mk]}$ equals $\sigma_{[ij]}$, the result in the proof of proposition 5 is applicable:

$$N(\log \frac{1}{N} SSE_{[mk]} - \log \frac{1}{N} SSE_{[ij]}) \to^d e^{W_{[mk]/[ij]}},$$

where $W_{[mk]/[ij]}$ has a non-degenerate distribution. In the other cases where the variances are different, the term with the variances dominates the term with the number of parameters.

Using this result, the differences in BIC values in the population are:

$$BIC_{[mk]} - BIC_{[ij]} = \begin{cases} \infty & \text{if } \sigma_{[mk]} > \sigma_{[ij]} \\ -\infty & \text{if } \sigma_{[mk]} < \sigma_{[ij]} \\ \infty & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} > k_{[ij]} \\ -\infty & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} < k_{[ij]} \\ e^{W_{[mk]/[ij]}} & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} = k_{[ij]} \end{cases}.$$

■

Proof: Proposition 10 (AIC and AICC). The choice of model can be determined by the differences in AIC. A model i is chosen over a model j if and only if

$$N(\log \frac{1}{N} SSE_i - \log \frac{1}{N} SSE_j) + 2(k_i - k_j) < 0$$

for all $i, j \neq m, k$.

The population equivalent or probability limit of $\frac{1}{N} SSE_{[mk]}$ is $\sigma^2_{[mk]}$. In case $\sigma_{[mk]}$ equals $\sigma_{[ij]}$, the result in the proof of proposition 3 can be applied to give:

$$N(\log \frac{1}{N} SSE_{[mk]} - \log \frac{1}{N} SSE_{[ij]}) \to^d e^{W_{[mk]/[ij]}},$$

where $W_{[mk]/[ij]}$ has a non-degenerate distribution. In the other cases where the variances are different, the term with the variances dominates the term with the number of parameters.

Using this result, the differences in AIC values in the population are:

$$AIC_{[mk]} - AIC_{[ij]} = \begin{cases} \infty & \text{if } \sigma_{[mk]} > \sigma_{[ij]} \\ -\infty & \text{if } \sigma_{[mk]} < \sigma_{[ij]} \\ e^{W_{[mk]/[ij]}} + 2(k_{[mk]} - k_{[ij]}) & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} > k_{[ij]} \\ e^{W_{[mk]/[ij]}} + 2(k_{[mk]} - k_{[ij]}) & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} < k_{[ij]} \\ e^{W_{[mk]/[ij]}} & \text{if } \sigma_{[mk]} = \sigma_{[ij]} \text{ and } k_{[mk]} = k_{[ij]} \end{cases} .$$

The corrected AIC is the same as AIC in the population since the correction term is 0 in the population. ■

Proof: Theorem 11 (Importance test). Assume that the regressors have mean 0. Let X_I be the regressors in I^m. Then the coefficients in the linear regression of $x_j (\notin I^m)$ on X_I are

$$\varphi = \left(E(X_I X_I')\right)^{-1} E(X_I x_j).$$

Assuming that $E(X_I X_I')$ has full rank, $\varphi = 0$ if and only if $E(X_I x_j) = 0$. That is, $E(x_j \mid X_I) = E(x_j)$ for $x_j \in (I^m)^c$ if and only if $corr(x_j, x_i) = 0$ for all $x_i \in I^m$. ■

References

[1] Anderson, D.R., Burnham, K.P., 2002. *Model Selection and Multimodel Inference: A practical information-Theoretic Approach.* Springer, USA.

[2] Barro, R., Sala-i-Martin, X., 2004. *Economic Growth.* MIT Press, USA.

[3] Burnside, C., Dollar, D., 2000. Aid, policies and Growth. *American Economic Review* 90(4), 847-68.

[4] Bleaney, M. and Nishiyama,A., 2002. Explaining Growth: A Contest Between Models. *Journal of Economic Growth* 7, 43-56.

[5] Breusch, T., 1986. Hypothesis Testing in Unidentified Models..*Review of Economic Studies* 53(4), 635-651.

[6] Breusch, T., 1990. Simplified Extreme Bounds. In *Modelling Economic Series.* Ed. C. Granger. Clarendon Press, Oxford.

[7] Carrasco, M., Florens, J., Renault, E., 2003. Linear Inverse Problems in Structural Econometrics Estimation based on Spectral Decomposition and regularization. forthcoming in *Handbook of Econometrics*, vol 6.

[8] Dalgaard, C., Hansen, H., Tarp, F., 2004. On the Empirics of Foreign Aid and growth. *The Economic Journal* 114(496), F244-71.

[9] Durlauf, S.N., Quah, D., 1999. The New Empirics of Economics Growth, *Handbook of Econometrics.* vol 1A. Edt. Taylor and Woodford, Elsevier.

[10] Durlauf, S.N., Johnson, P., Temple, J., Forthcoming. Growth Econometrics, *Handbook of Economic Growth.* Edt. Aghion, P. and S. N. Durlauf, Elsevier.

[11] Easterly, W., Levine, R., Roodman, D., 2004. Aid, policies, and growth: Comment. *American Economic Review* 94(3), 774-780.

[12] Fernandez, C. Ley, E. and Steele, M.F.J., 2001a. Model Uncertainty in Cross-country Growth regressions. *Journal of applied Econometrics* 16(5), 563-576.

[13] Fernandez,C. , Ley, E. and Steele, M.F.J., 2001b. Benchmark priors for Bayesian Model averaging. *Journal of Econometrics* 100(2), 381-472.

[14] Granger, C., Uhlig, H., 1990. Reasonable extreme bound analysis. *Journal of Econometrics* 44, 159-170.

[15] Goldberger, A.S., 1998. *Introductory Econometrics.* Harvard University Press.

[16] Haan, J., Sturm, J-E., 2000. On the relationship between economic freedom and economic growth. *European Journal of Political Economy* 16(2), 215-241.

[17] Hansen, B.E., 1999. Discussion of 'Data mining reconsidered., *Econometrics Journal* 2, 192-201.

[18] Hansen, P.R., 2003. Regression Analysis with Many Specifications: A Boostrap Method for Robust Inference. Working Paper.

[19] Hendry, D. F. 1995. *Dynamic Econometrics.* Oxford: Oxford University Press.

[20] Hendry, D.F., Krolzig, M., 1999. Improving on 'Data mining reconsiderd' by K.D. Hoover and S.J.Perez. *Econometrics Journal* 2, 202-219.

[21] Hendry, D.F., Krolzig, M., 2004. We ran one regression. *Oxford Bulletin of Economics and Statistics* 66(5), 799-810.

[22] Hoover, K., Perez, K., 1999. Data-mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal* 2, 167-191.

[23] Hoover, K., Perez, K., 2004. Truth and Robustness in Cross-country Growth Regressions. *Oxford Bulletin of Economics and Statistics*, 66(5), 765-798.

[24] Kalaitzidakis, P., Mamuneas, T.P., Stengos,T., 2002. Specification and sensitivity analysis of cross-country growth regressions. *Empirical Economics* 27, 645-656.

[25] Leamer, E., 1983. Let's take the con out of econometrics. *American Economic Review* 73(1), 31-43.

[26] Levine, R., Renelt, D., 1992. A Sensitivity Analysis of Cross-Country Growth Regressions. *American Economic Review* 82(2), 942-963

[27] McAleer, M., Pagan, A.R., Volker, P.A., 1985. What Will Take the Con Out of Econometrics? *The American Economic Review* 75(3), 293-307.

[28] McAleer, M., 1994. Sherlock Holmes and the Search for Truth: A Diagnostic Tale. *Journal of Economic Surveys* 8(4), 317-370.

[29] Mittelhammer, R.C., Judge. G.G., Miller, D.J., 2000. *Econometric Foundations.* Cambridge University Press, USA.

[30] Pagan, A., 1987. Three Econometric Methodologies: A critical Appraisal. *Journal of Economic Surveys* 1(1), 3-24

[31] Rao, C. R., 1973. *Linear Statistical Inference and Its Applications.* Second edition. John Wiley & Sons, Inc, USA.

[32] Roodman, D., 2004. The Anarchy of Numbers:Aid, Development, and Cross-country Empirics, Working paper, Center for Global Development.

[33] Sala-i-Martin, X., 1997a. I Just Ran Four Million Regressions. NBER Working paper no.w6252.

[34] Sala-i-Martin, X., 1997b. I Just Ran Two Million Regressions. *American Economic Review* 87(2) , 178-183.

[35] Sala-i-Martin, X., Doppelhoffer, G., Miller, R., 2004. Determinants of Long-Term growth: A Bayesian Averaging of Classical Estimates (BACE) approach. *American Economic Review* 94(4), 813-835.

[36] Scheffé, H., 1959. *The Analysis of Variance.* John Wiley and Sons, Inc, New York.

[37] Stock, J. and Watson, M., 2002. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association* 97(460), 1180-1191.

[38] Sturm, J-E., de Haan, J., forthcoming. Determinants of Long-term Growth: New Results Applying Robust Estimation and Extreme Bounds Analysis. *Empirical Economics.*

[39] Temple, J., 2000. Growth regressions and what the textbooks don't tell you. *Bulletin of Economic Research* 53(3), 181-205.

[40] White, H., 1999. A reality check for data snooping. *Econometrica,* 68, 1097-1127.

[41] Wooldridge, J., 2002. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, USA.