Journal of Econometrics 154 (2010) 139–153

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A comparison of two model averaging techniques with an application to growth empirics

Jan R. Magnus[a,*], Owen Powell[b], Patricia Prüfer[c]

[a] Department of Econometrics & Operations Research, Tilburg University, The Netherlands

[b] CentER, Tilburg University, The Netherlands

[c] Netherlands Bureau for Economic Policy Analysis (CPB) and CentER, Tilburg University, The Netherlands

ARTICLE INFO

Article history:
Received 5 April 2008
Received in revised form
18 May 2009
Accepted 25 July 2009
Available online 6 August 2009

JEL classification:
C51
C52
C13
C11

Keywords:
Model averaging
Bayesian analysis
Growth determinants

ABSTRACT

Parameter estimation under model uncertainty is a difficult and fundamental issue in econometrics. This paper compares the performance of various model averaging techniques. In particular, it contrasts Bayesian model averaging (BMA) — currently one of the standard methods used in growth empirics — with a new method called weighted-average least squares (WALS). The new method has two major advantages over BMA: its computational burden is trivial and it is based on a transparent definition of prior ignorance. The theory is applied to and sheds new light on growth empirics where a high degree of model uncertainty is typically present.

1. Introduction

This paper has two purposes. First, it introduces a new model averaging technique, called weighted-average least squares (hereafter WALS), which we claim to be theoretically and practically superior to standard Bayesian model averaging (BMA). It is theoretically superior because it treats our ignorance about the priors in a different manner, thereby obtaining a better risk profile and, in particular, avoiding unbounded risk. It is practically superior because the space over which we need to perform model selection increases *linearly* rather than exponentially in size. Thus, if we have sixty regressors to search over (which is not unusual in the growth literature), then computing time of standard BMA is of the order 2^{60}, while computing time of WALS is of the order 60. This means that what WALS can do in one second, BMA can only do in six hundred million years. Exact computation of a complete BMA is therefore rarely done; instead some Markov chain Monte Carlo (MCMC) method is typically applied.

The second purpose is to contribute to the debate on growth empirics. Since the seminal studies of Kormendi and Meguire (1985) and Barro (1991), empirical research on the determinants of economic growth has identified numerous variables as being robustly (partially) correlated with productivity growth in an economy. Durlauf et al. (2005) list 145 potential right-hand side variables for growth regressions and cluster them into more than forty areas (or theories), such as human capital, finance, government, and trade. Taking into account the limited number of observations available at a national level, growth empirics has been heavily criticized because of the inherent model uncertainty; see Durlauf et al. (2005) for a recent in-depth survey.

Sometimes growth theory can support choices of specific variables, but the inclusion or exclusion of most variables is typically arbitrary, a phenomenon labeled the 'open-endedness' of growth theory (Brock and Durlauf, 2001). In addition, while theory may provide general qualitative variables (such as human capital), it does not tell us how these variables are to be specified or measured. We are thus faced with (at least) two types of uncertainty, each of which brings about model uncertainty. Since there exist a wide set of possible model specifications, we often obtain contradictory conclusions. To make matters worse, estimation results are often not robust to small changes in model specification, making credible interpretations of the results

* Corresponding address: Department of Econometrics & OR, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. Tel.: +31 13 466 3092.
E-mail address: magnus@uvt.nl (J.R. Magnus).

0304-4076/$ – see front matter
doi:10.1016/j.jeconom.2009.07.004

hazardous. A proper treatment of model uncertainty is clearly important.

One such treatment is model averaging, where the aim of the investigator is not to find the best possible model, but rather to find the best possible estimates. Each model contributes information about the parameters of interest, and all these pieces of information are combined taking into account the trust we have in each model, based on our prior beliefs and on the data.

In a sense, *all* estimation procedures are model averaging algorithms, although possibly extreme or limiting cases. Our framework is the linear regression model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon = X\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2), \tag{1}$$

where y $(n \times 1)$ is the vector of observations, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are matrices of nonrandom regressors, ε is a random vector of unobservable disturbances, and β_1 and β_2 are unknown parameter vectors. We assume that $k_1 \geq 1, k_2 \geq 0, k := k_1 + k_2 \leq n - 1$, that $X := (X_1 : X_2)$ has full column-rank, and that the disturbances $(\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d. $\mathrm{N}(0, \sigma^2)$.

The reason for distinguishing between X_1 and X_2 is that X_1 contains explanatory variables which we want in the model on theoretical or other grounds (irrespective of the found t-ratios of the β_1-parameters), while X_2 contains additional explanatory variables of which we are less certain. The columns of X_1 are called 'focus' regressors, and the columns of X_2 'auxiliary' regressors. [1]

There are k_2 components of β_2, and a different model arises whenever a different subset of the β_2's is set equal to zero. If $k_2 = 0$, then no model selection takes place. If $k_2 = 1$, then there are two models to consider: the unrestricted and the restricted model. If $k_2 = 2$, there are four models: the unrestricted, two partially restricted (where one of the two β_2's is zero), and the restricted model. In general, there are 2^{k_2} models to consider. We denote the ith model by $\mathcal{M}_i$, which we write as

$$y = X_1\beta_1 + X_{2i}\beta_{2i} + \varepsilon,$$

where X_{2i} denotes an $n \times k_{2i}$ matrix containing a subset of k_{2i} columns of X_2, and β_{2i} denotes the corresponding $k_{2i} \times 1$ subvector of β_2. We have of course $0 \leq k_{2i} \leq k_2$.

Model averaging estimation proceeds in two steps. In the first step we ask how to estimate the parameters, conditional upon a selected model. In the second step we compute the estimator as a weighted average of these conditional estimators. There exist both Bayesian and non-Bayesian ideas about how to estimate and how to find the weights. Our emphasis will be on the Bayesian framework; for the non-Bayesian approach, see Claeskens and Hjort (2003), Hjort and Claeskens (2003), Hansen (2007), and Liang et al. (2008).

The *unrestricted* estimator simply sets the weight for the unrestricted model (no restrictions on β_1 or β_2) to one and performs a single estimation. Similarly, the *restricted* estimator sets β_2 to zero and estimates the resulting restricted model. Both estimators are, admittedly trivial, examples of a model averaging procedure. More interesting, and more common, are *general-to-specific* (GtS) estimators which do involve a model selection procedure, typically based on the 'significance' of parameters through their t-ratios. There are many problems with this procedure (see e.g. Magnus, 1999), but the most important is that the model selection procedure is completely separated from the estimation procedure. What is reported are therefore conditional estimates, but the researcher acts as if they are unconditional estimates. This problem is known as *pretesting*.

In order to combine model selection and estimation, the Bayesian method offers a natural framework. The basic equations of BMA were first presented by Leamer (1978, Sections 4.4–4.6), who proposed Bayesian averaging of *Bayesian* estimates. In the context of growth econometrics, BMA was first applied by Fernández et al. (2001a) and Brock and Durlauf (2001). BMA is flexible with respect to the size and exact specification of a model and it does not require the a priori selection of any model. Inference is based on a weighted average over all models. The idea of Bayesian averaging of *classical* estimates was first proposed in Raftery (1995) and later by Sala-i-Martin et al. (2004). In growth econometrics, BMA has proved useful, and recent applications include León-González and Montolio (2004), Sala-i-Martin et al. (2004) and Masanjala and Papageorgiou (2007). Recently, interest is growing in different aspects of growth empirics, such as nonlinearities, parameter heterogeneity, and endogeneity. BMA is also applied in other areas of economics; see for example Tsangarides et al. (2004), Crespo-Cuaresma and Doppelhofer (2007), Eicher et al. (2007a,c), Masanjala and Papageorgiou (2008) and Prüfer and Tondl (2008). In short, BMA has become an important technique.

There are, however, two major problems with BMA. First, the computational burden is very substantial. In fact, it is usually impossible to get exact BMA estimates, in which case some MCMC method must be applied, of which the Metropolis–Hastings algorithm is the most common. Second, Bayesian techniques work well when prior information is available, in which case they guide us as to how this information should be combined with information from the data. But when no prior information is available and nevertheless informative priors need to be specified (as is the case with BMA), then we need to reflect on the meaning and impact of these priors.

In addition to these two problems, there are some further uncomfortable aspects to BMA. One is that BMA takes different priors for the same parameter depending on which submodel is considered. This was also noted by Hjort and Claeskens (2003), and it is a little difficult to interpret. Another uncomfortable aspect is that — since exact BMA is computationally so demanding — it is very difficult to consider extensions, for example to nonspherical disturbances.[2]

Our proposed WALS method deals with all these problems. The computational burden is trivial, and the proposed prior is attractive because it is 'neutral' (mimicking ignorance) and also near-optimal in the sense of minimizing some risk or regret criterion (Magnus, 2002). It is based on the equivalence theorem of Magnus and Durbin (1999) and Danilov and Magnus (2004), and was originally developed to better understand pretesting.

The concept and treatment of ignorance is essential in both BMA and WALS. Suppose for simplicity that $k_2 = 1$ in (1), so that there is only one auxiliary regressor x_2 and only one auxiliary parameter β_2, and we have $y = X_1\beta_1 + \beta_2 x_2 + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. It is well-known that if we delete the auxiliary variable x_2 from our regression equation, then R^2 will always decrease, but $\bar{R}^2$ (the adjusted R^2) will decrease if, and only if, the t-ratio of the auxiliary

[1] Maybe Leamer (1978, p. 194) was the first to categorize variables into two classes, which he called 'focus' and 'doubtful', so that the focus variables are always in the model, while the doubtful variables can be combined in an arbitrary linear manner, a special case of which is exclusion. Later Leamer (1985, p. 309) preferred the use of 'free' instead of 'focus', because it is not always the case that the focus variables are the focus of a study — they are just the variables that are always in the equation.

[2] See, however, Doppelhofer and Weeks (2008) who study the robustness of BMA with respect to outliers and heteroskedasticity in the context of cross-country growth regressions. Magnus et al. (2009) extend WALS estimation to nonspherical disturbances.

parameter is smaller than one in absolute value. It is also well-known (Magnus and Durbin, 1999, Theorem 1) that if we define the 'theoretical' t-ratio

$$\eta := \frac{\beta_2}{\sigma/\sqrt{x_2' M_1 x_2}}, \qquad M_1 := I_n - X_1(X_1'X_1)^{-1}X_1',$$

then $\text{MSE}(\hat{\beta}_{1r}) \leq \text{MSE}(\hat{\beta}_{1u})$ if, and only if, $|\eta| \leq 1$, where $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ denote the restricted (with $\beta_2 = 0$) and unrestricted estimator of β_1 respectively. Hence we shall say that we are 'ignorant' (or 'neutral') about the auxiliary parameter β_2 when (a) we do not know whether β_2 is positive or negative, and (b) we do not know whether including the corresponding auxiliary regressor x_2 will increase or decrease the mean squared error of the estimated focus parameter β_1. More formally, we choose the prior distribution in WALS such that the prior median of η is zero and the prior median of η^2 is one. This treatment of ignorance is further elaborated on and defended in Magnus (2002), and it is close to the idea in Masanjala and Papageorgiou (2008), who state that a posterior inclusion probability of 0.50 corresponds approximately to an absolute t-ratio of one. The proposed priors for WALS are taken from the Laplace distribution and thus generate bounded risk, in contrast to the normal prior adopted by BMA which generates unbounded risk. Fig. 1 in Section 3.4 illustrates this essential difference.

In this paper we confront BMA with WALS, and apply both techniques to shed further light on the determinants of economic growth. In our growth estimations we use a set-up which allows us to distinguish between standard Solow growth determinants and determinants that have been suggested in the so-called 'new growth' theories. Based on these analyses, we can not only draw conclusions on the most appropriate model averaging technique but also provide insights on the impacts of frequently used growth determinants.

The paper is organized as follows. The two main model averaging techniques are described in Sections 2 (BMA) and 3 (WALS). In Section 2 we extend the standard BMA theory to allow for the case where model selection takes place over a subset of the regressors. In Section 3 we extend the theory of WALS (developed in the context of pretesting), so that it can be used as a general model averaging technique. Sections 4–6 present the growth estimation set-up and results, and Section 7 concludes. An Appendix contains a description and justification of our data and selected variables.

2. Bayesian model averaging (BMA)

The usual set-up for Bayesian model averaging is the special case of (1) where $k_1 = 1$ and $X_1 = \imath$ (the vector of ones), so that the constant term is present in all models and model selection takes place over all regressors except the constant term. Our treatment is more general and allows model selection to take place over a subset (X_2) of the regressors, while the focus regressors (the columns of X_1) are forced to be present in every model.

A very large literature exists on BMA, some of which is mentioned in the introduction. Useful literature summaries can be found in Raftery et al. (1997) and Hoeting et al. (1999).

2.1. Prior, likelihood, and posterior in model $\mathcal{M}_i$

Assuming that $\mathcal{M}_i$ is the true model, the likelihood is given by

$$p(y \mid \beta_1, \beta_{2i}, \sigma^2, \mathcal{M}_i) \propto (\sigma^2)^{-n/2} \exp -\frac{S_i}{2\sigma^2}, \tag{2}$$

where $S_i := (y - X_1\beta_1 - X_{2i}\beta_{2i})'(y - X_1\beta_1 - X_{2i}\beta_{2i})$. Following standard Bayesian theory of the normal linear model (O'Hagan, 1994, Chapter 9), we impose the conventional improper prior distribution $p(\sigma^2|\mathcal{M}_i) \propto \sigma^{-2}$ together with a partially proper prior on $\beta_1, \beta_{2i}|\sigma^2, \mathcal{M}_i$:

$$p(\beta_1 \mid \sigma^2, \mathcal{M}_i) \propto 1, \qquad \beta_{2i} \mid \beta_1, \sigma^2, \mathcal{M}_i \sim \text{N}(0, \sigma^2 V_{0i}),$$

where V_{0i} is a positive definite $k_{2i} \times k_{2i}$ matrix to be specified later. The joint prior distribution is then

$$p(\beta_1, \beta_{2i}, \sigma^2 \mid \mathcal{M}_i) \propto (\sigma^2)^{-(k_{2i}+2)/2} \exp -\frac{\beta_{2i}' V_{0i}^{-1} \beta_{2i}}{2\sigma^2}. \tag{3}$$

To deal with partially proper (informative), partially improper (noninformative) priors is no trivial matter; see Bauwens et al. (1999, pp. 117–118). Our approach will be to think of the improper prior distribution as a special case of the following *proper* prior distribution:

$$p(\beta_1, \beta_{2i}, \sigma^2 \mid \mathcal{M}_i) \propto (\sigma^2)^{-(d_0+k_1+k_{2i}+2)/2} \times \exp -\frac{h_0\beta_1'\beta_1 + \beta_{2i}'V_{0i}^{-1}\beta_{2i} + a_0}{2\sigma^2}, \tag{4}$$

where the special case (3) occurs when $h_0 = 0$, $a_0 = 0$, and $d_0 = -k_1$.[3]
Combining the prior (4) with the likelihood (2) gives the posterior

$$p(\beta_1, \beta_{2i}, \sigma^2 \mid y, \mathcal{M}_i) \propto (\sigma^2)^{-(d+k_1+k_{2i}+2)/2} \exp -\frac{R_i + a_i}{2\sigma^2}, \tag{5}$$

where $d = d_0 + n$,

$$R_i := \begin{pmatrix} \beta_1 - b_{1i} \\ \beta_{2i} - b_{2i} \end{pmatrix}' V_i^{-1} \begin{pmatrix} \beta_1 - b_{1i} \\ \beta_{2i} - b_{2i} \end{pmatrix},$$

$$V_i^{-1} := \begin{pmatrix} X_1'X_1 + h_0 I_{k_1} & X_1'X_{2i} \\ X_{2i}'X_1 & X_{2i}'X_{2i} + V_{0i}^{-1} \end{pmatrix},$$

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} := V_i(X_1 : X_{2i})'y,$$

and

$$a_i := a_0 + y'y - y'(X_1 : X_{2i})V_i(X_1 : X_{2i})'y.$$

Hence the posterior density of β_1, β_{2i}, and σ^2 — given the data y and model $\mathcal{M}_i$ — is the familiar normal-inverse-gamma distribution with parameters a_i, d, (b_{1i}, b_{2i}), and V_i.

A little algebra gives

$$V_i = \begin{pmatrix} (X_1'X_1 + h_0 I_{k_1})^{-1} + Q_i V_{2i} Q_i' & -Q_i V_{2i} \\ -V_{2i} Q_i' & V_{2i} \end{pmatrix}, \tag{6}$$

where

$$V_{2i}^{-1} = V_{0i}^{-1} + X_{2i}' M_1^* X_{2i}, \qquad M_1^* = I_n - X_1(X_1'X_1 + h_0 I_{k_1})^{-1} X_1',$$

and

$$Q_i = (X_1'X_1 + h_0 I_{k_1})^{-1} X_1' X_{2i}.$$

From (6) we find

$$(X_1 : X_{2i})V_i(X_1 : X_{2i})' = I - M_1^* + M_1^* X_{2i} V_{2i} X_{2i}' M_1^*,$$

so that we can rewrite a_i as

$$a_i = a_0 + y'y - y'(X_1 : X_{2i})V_i(X_1 : X_{2i})'y = a_0 + y'(M_1^* - M_1^* X_{2i} V_{2i} X_{2i}' M_1^*)y.$$

[3] Whenever priors are used the question of sensitivity of the posterior moments to the priors is important. We do not examine this issue here. Recent examples of such prior robustness checks for BMA include Ley and Steel (2009) and Eicher et al. (2007b).

We now specialize to the improper prior given in (3) by setting $h_0 = 0$, $a_0 = 0$, and $d_0 = -k_1$. The matrix M_1^* then specializes to the idempotent matrix $M_1 := I_n - X_1(X_1'X_1)^{-1}X_1'$, and we have $a_i = (M_1y)'A_i(M_1y)$, where $A_i := M_1 - M_1X_{2i}V_{2i}X_{2i}'M_1$. We notice that a_i is a function of M_1y only and does not depend on $X_1'y$. It follows that

$$\mathrm{E}(\beta_1 \mid y, \mathcal{M}_i) = b_{1i} = (X_1'X_1)^{-1}X_1'(y - X_{2i}b_{2i}), \tag{7}$$

$$\mathrm{E}(\beta_{2i} \mid y, \mathcal{M}_i) = b_{2i} = (V_{0i}^{-1} + X_{2i}'M_1X_{2i})^{-1}X_{2i}'M_1y, \tag{8}$$

and, when $n > k_1 + 2$,

$$\mathrm{var}(\beta_1 \mid y, \mathcal{M}_i) = \frac{a_i}{n - k_1 - 2}((X_1'X_1)^{-1} + Q_iV_{2i}Q_i'), \tag{9}$$

$$\mathrm{var}(\beta_{2i} \mid y, \mathcal{M}_i) = \frac{a_i}{n - k_1 - 2}V_{2i}. \tag{10}$$

2.2. Marginal likelihood of model $\mathcal{M}_i$

In order to find the marginal likelihood we return to the proper prior (4). Since $|V_i| = |X_1'X_1 + h_0I_{k_1}|^{-1} \cdot |V_{2i}|$, we obtain the marginal density of y in model $\mathcal{M}_i$ as

$$\begin{aligned} p(y \mid \mathcal{M}_i) &= \int\int\int p(y \mid \beta_1, \beta_{2i}, \sigma^2, \mathcal{M}_i)\, p(\beta_1, \beta_{2i}, \sigma^2 \mid \mathcal{M}_i) \\ &\quad \times \mathrm{d}\beta_1\, \mathrm{d}\beta_{2i}\, \mathrm{d}\sigma^2 \\ &= \frac{\pi^{-n/2} h_0^{k_1/2} a_0^{d_0/2}\, \Gamma(d/2)}{|X_1'X_1 + h_0I_{k_1}|^{1/2}\, \Gamma(d_0/2)} \cdot \frac{|V_{0i}^{-1}|^{1/2}}{|V_{2i}^{-1}|^{1/2}} \cdot a_i^{-d/2} \\ &= c \cdot |V_{2i}^{-1}|^{-1/2} |V_{0i}^{-1}|^{1/2} a_i^{-d/2}, \end{aligned}$$

where c is a normalizing constant which does not depend on i or y.

Now specializing to the improper prior (3) by setting $h_0 = 0$, $a_0 = 0$, and $d_0 = -k_1$, we find

$$p(y \mid \mathcal{M}_i) = c \cdot \frac{|V_{0i}^{-1}|^{1/2}}{|V_{0i}^{-1} + X_{2i}'M_1X_{2i}|^{1/2}} \cdot (y'M_1A_iM_1y)^{-(n-k_1)/2}, \tag{11}$$

where

$$A_i := M_1 - M_1X_{2i}\left(V_{0i}^{-1} + X_{2i}'M_1X_{2i}\right)^{-1}X_{2i}'M_1$$

and $M_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$. If we let $p(\mathcal{M}_i)$ denote the prior probability that $\mathcal{M}_i$ is the true model, and $\lambda_i := p(\mathcal{M}_i|y)$ the posterior probability for model $\mathcal{M}_i$, then

$$\lambda_i = \frac{p(\mathcal{M}_i)p(y \mid \mathcal{M}_i)}{\sum_j p(\mathcal{M}_j)p(y \mid \mathcal{M}_j)} \quad (i = 1, \ldots, 2^{k_2}),$$

We shall assign equal prior probability to each model under consideration. This seems to be in line with the standard literature on BMA, although it is not without criticism and alternative choices for $p(\mathcal{M}_i)$ have been proposed. Many researchers feel that simpler models should be preferred to more complex ones, all else being equal. Durlauf et al. (2005), on the other hand, find the idea of promoting parsimonious models through the priors unappealing. Brock and Durlauf (2001) raise objections against uniform priors on the model space because of the implicit assumption that the probability that one regressor appears in the model is independent of the inclusion of others, whereas, in fact, regressors are typically correlated. They suggest a hierarchical structure for the model prior. This, however, requires agreement on which regressors are proxies for the same theories. As stated in Eicher et al. (2007b), such an agreement is usually not within reach and, therefore, independent model priors seem a reasonable compromise. Thus motivated we write

$$p(\mathcal{M}_i) = 2^{-k_2}.$$

Then $\lambda_i = p(y \mid \mathcal{M}_i)$, where the normalizing constant c is chosen such that $\sum_i \lambda_i = 1$.

2.3. Model averaging

So far we have conditioned on one model, namely model $\mathcal{M}_i$. In the Bayesian framework it is now easy to consider all models in our assumed model space $\mathcal{M} := \{\mathcal{M}_i,\ i = 1, \ldots, 2^{k_2}\}$, by writing the posterior distribution of our parameters β_1, β_2, and σ^2 given the data y as

$$p(\beta_1, \beta_2, \sigma^2 \mid y) = \sum_{i=1}^{2^{k_2}} \lambda_i\, p(\beta_1, \beta_{2i}, \sigma^2 \mid y, \mathcal{M}_i). \tag{12}$$

This is a weighted average of the posterior distributions under each model, weighted by the corresponding posterior model probabilities.

The posterior mean and variance of β_1 are

$$b_1 := \mathrm{E}(\beta_1 \mid y) = \sum_i \lambda_i b_{1i}, \tag{13}$$

and

$$\mathrm{var}(\beta_1 \mid y) = \sum_i \lambda_i \left(V_{1i}^* + b_{1i}b_{1i}'\right) - b_1b_1', \tag{14}$$

where $b_{1i} := \mathrm{E}(\beta_1 \mid y, \mathcal{M}_i)$ and $V_{1i}^* := \mathrm{var}(\beta_1 \mid y, \mathcal{M}_i)$; see Raftery (1993) and Draper (1995).

To obtain the corresponding results for β_2 we introduce the $k_2 \times k_{2i}$ selection matrices T_i with full column-rank, so that $T_i' = (I_{k_{2i}} : 0)$ or a column-permutation thereof, and $T_i\beta_{2i}$ is the $k_2 \times 1$ vector obtained from β_2 by setting the components not included in $\mathcal{M}_i$ to zero. The posterior mean and variance of β_2 are then

$$b_2 := \mathrm{E}(\beta_2 \mid y) = \sum_i \lambda_i T_i b_{2i}, \tag{15}$$

and

$$\mathrm{var}(\beta_2 \mid y) = \sum_i \lambda_i T_i \left(V_{2i}^* + b_{2i}b_{2i}'\right) T_i' - b_2b_2', \tag{16}$$

where $b_{2i} := \mathrm{E}(\beta_{2i} \mid y, \mathcal{M}_i)$ and $V_{2i}^* := \mathrm{var}(\beta_{2i} \mid y, \mathcal{M}_i)$.

2.4. Implementation using g-priors

Following Zellner (1986) we assume that the prior variance V_{0i} is given by

$$V_{0i}^{-1} = g_iX_{2i}'M_1X_{2i} \quad (g_i > 0).$$

This gives

$$\lambda_i = c \cdot \left(\frac{g_i}{1 + g_i}\right)^{k_{2i}/2} (y'M_1A_iM_1y)^{-(n-k_1)/2},$$

where

$$A_i = \frac{g_i}{1 + g_i}M_1 + \frac{1}{1 + g_i}(M_1 - M_1X_{2i}(X_{2i}'M_1X_{2i})^{-1}X_{2i}'M_1).$$

We have

$$b_{1i} = (X_1'X_1)^{-1}X_1'(y - X_{2i}b_{2i}),$$

$$b_{2i} = \frac{1}{1 + g_i}(X_{2i}'M_1X_{2i})^{-1}X_{2i}'M_1y.$$

Also, when $n > k_1 + 2$ and defining $s_i^2 := y'M_1A_iM_1y/(n - k_1 - 2)$, we find

$$V_{1i}^* = s_i^2(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'X_{2i}V_{2i}^*X_{2i}'X_1(X_1'X_1)^{-1},$$

$$V_{2i}^* = \frac{s_i^2}{1 + g_i}(X_{2i}'M_1X_{2i})^{-1}.$$

Our final ingredient is the specification of g_i. We follow Fernández et al. (2001b) and choose

$$g_i := \frac{1}{\max(n, k_2^2)},$$

where we note that g_i is the same for all i. One alternative would have been $g_i := 1/n$, the so-called 'unit information prior' (Raftery, 1995), recently advocated by Eicher et al. (2007b) and Masanjala and Papageorgiou (2008). In our case with $n = 74$ and $k_2 = 4, 9$, or 12, the difference between the two priors is negligible.

The above results now allow us to calculate the BMA estimates and precisions of β_1 and β_2 from (13)–(16). Special cases arise and some care is required when $k_2 = 0$ (no model selection) or $k_1 = 0$ (model selection takes place over all regressors). Our Matlab program, downloadable from http://center.uvt.nl/staff/magnus/wals, allows for these special cases.

3. Weighted-average least squares (WALS)

3.1. Orthogonalization

Weighted-average least-squares estimation starts with the realization that we can 'orthogonalize' the columns of X_2 such that $X_2'M_1X_2 = I_{k_2}$, where we recall that $M_1 := I_n - X_1(X_1'X_1)^{-1}X_1'$. More precisely, if we let P be an orthogonal $k_2 \times k_2$ matrix such that $P'X_2'M_1X_2P = \Lambda$ (diagonal), and define new auxiliary regressors $X_2^* := X_2P\Lambda^{-1/2}$ and new auxiliary parameters $\beta_2^* = \Lambda^{1/2}P'\beta_2$, then $X_2^*\beta_2^* = X_2\beta_2$ and $X_2^{*\prime}M_1X_2^* = I_{k_2}$. There are major advantages in working with X_2^* and β_2^* instead of X_2 and β_2, as will become clear shortly. Hence, we shall *initially* assume that this orthogonalization has taken place.

Assumption 1. $X_2'M_1X_2 = I_{k_2}$.

Assumption 1 thus requires that the columns $x_{21}, \ldots, x_{2k_2}$ of X_2 (the auxiliary regressors) are 'orthogonal' in the sense that M_1x_{2i} and M_1x_{2j} are orthogonal for every $i \neq j$. This will not affect the interpretation of the β_1-coefficients, but it will change the interpretation of the β_2-coefficients. However, we can always recover β_2 from $\beta_2 = P\Lambda^{-1/2}\beta_2^*$.

3.2. Restricted least squares

Given Assumption 1, the least-squares (LS) estimators of β_1 and β_2 in the unrestricted model (1) are

$$\hat{\beta}_1 = \hat{\beta}_{1r} - Q\hat{\beta}_2, \qquad \hat{\beta}_2 = X_2'M_1y,$$

where $\hat{\beta}_{1r} := (X_1'X_1)^{-1}X_1'y$ and $Q := (X_1'X_1)^{-1}X_1'X_2$. The subscript 'r' denotes 'restricted' (with $\beta_2 = 0$). We see that $\hat{\beta}_2 \sim \mathrm{N}(\beta_2, \sigma^2 I_{k_2})$.

Let S_i be an $k_2 \times (k_2 - k_{2i})$ selection matrix with full column-rank, where $0 \leq k_{2i} \leq k_2$, so that $S_i' = (I_{k_2-k_{2i}} : 0)$ or a column-permutation thereof. We are interested in the restricted LS estimators of β_1 and β_2, the restriction being $S_i'\beta_2 = 0$. Let $\mathcal{M}_i$ denote the linear model (1) under the restriction $S_i'\beta_2 = 0$, and denote the LS estimators of β_1 and β_2 in model $\mathcal{M}_i$ by $\hat{\beta}_{1i}$ and $\hat{\beta}_{2i}$. Following Danilov and Magnus (2004, Lemmas A1 an A2), the restricted LS estimators of β_1 and β_2 are given by

$$\hat{\beta}_{1i} = \hat{\beta}_{1r} - QW_i\hat{\beta}_2, \qquad \hat{\beta}_{2i} = W_i\hat{\beta}_2, \tag{17}$$

where $W_i := I_{k_2} - S_iS_i'$ is a *diagonal* $k_2 \times k_2$ matrix with k_{2i} ones and $(k_2 - k_{2i})$ zeros on the diagonal, such that the jth diagonal element of W_i is zero if β_{2j} is restricted to be zero, and one otherwise. (If $k_{2i} = k_2$ then $W_i := I_{k_2}$.) The joint distribution of $\hat{\beta}_{1i}$ and $\hat{\beta}_{2i}$ is then

$$\begin{pmatrix}\hat{\beta}_{1i}\\ \hat{\beta}_{2i}\end{pmatrix} \sim \mathrm{N}_k\left(\begin{pmatrix}\beta_1 + QS_iS_i'\beta_2\\ W_i\beta_2\end{pmatrix}, \sigma^2\begin{pmatrix}(X_1'X_1)^{-1} + QW_iQ' & -QW_i\\ -W_iQ' & W_i\end{pmatrix}\right),$$

the residual vector $e_i := y - X_1\hat{\beta}_{1i} - X_2\hat{\beta}_{2i}$ is given by $e_i = D_iy$, where $D_i := M_1 - M_1X_2W_iX_2'M_1$ is a symmetric idempotent matrix of rank $n - k_1 - k_{2i}$, and the distribution of $s_i^2 := e_i'e_i/(n - k_1 - k_{2i})$ is

$$\frac{(n - k_1 - k_{2i})s_i^2}{\sigma^2} \sim \chi^2\left(n - k_1 - k_{2i}, \frac{\beta_2'S_iS_i'\beta_2}{\sigma^2}\right).$$

It follows that:

- all models which include x_{2j} as a regressor will have the *same* estimator of β_{2j}, namely $\hat{\beta}_{2j}$, irrespective which other β_2's are estimated;
- the estimators $\hat{\beta}_{21}, \hat{\beta}_{22}, \ldots, \hat{\beta}_{2k_2}$ are independent;
- if σ^2 is known or is estimated by s^2 (the LS estimator in the unrestricted model), then all models which include x_{2j} as a regressor yield the same t-ratio of β_{2j}.

3.3. The equivalence theorem

We now define the WALS estimator of β_1 as

$$b_1 = \sum_{i=1}^{2^{k_2}} \lambda_i\hat{\beta}_{1i}, \tag{18}$$

where the sum is taken over all 2^{k_2} different models obtained by setting a subset of the β_2's equal to zero, and the λ_i are weight-functions satisfying certain minimal regularity conditions, namely

$$\lambda_i \geq 0, \qquad \sum_i \lambda_i = 1, \quad \lambda_i = \lambda_i(M_1y). \tag{19}$$

The WALS estimator can then be written as $b_1 = \hat{\beta}_{1r} - QW\hat{\beta}_2$, where $W := \sum_i \lambda_iW_i$. Notice that, while the W_i are nonrandom, W is random. For example, when $k_2 = 2$, we have four models to compare: the restricted $\mathcal{M}_0$ ($\beta_{21} = \beta_{22} = 0$), the partially restricted $\mathcal{M}_1$ ($\beta_{22} = 0$) and $\mathcal{M}_2$ ($\beta_{21} = 0$), and the unrestricted $\mathcal{M}_{12}$. The corresponding W_i are

$$W_0 = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}, \qquad W_1 = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, \qquad W_2 = \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}, \qquad W_{12} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},$$

and hence

$$W = \begin{pmatrix}\lambda_1 + \lambda_{12} & 0\\ 0 & \lambda_2 + \lambda_{12}\end{pmatrix},$$

where the weight-functions λ_i are now labeled $\lambda_0, \lambda_1, \lambda_2, \lambda_{12}$, corresponding to the four models $\mathcal{M}_i$ and matrices W_i. We see that λ_0 does not appear in the matrix W and that W is diagonal because of Assumption 1.

A few words about the regularity conditions are in order. If σ^2 is known, then most or all diagnostics will use statistics (such as t- and F-statistics) which depend on $\hat{\beta}_2$ only. If σ^2 is not known and estimated by s^2, then all t- and F-statistics will depend on $(\hat{\beta}_2, s^2)$. Now, it is a basic result in least-squares theory that s^2 is independent of $(\hat{\beta}_1, \hat{\beta}_2)$. It follows that $\hat{\beta}_{1r}$ is independent of s^2. Hence, $\hat{\beta}_{1r}$

will be independent of $(\hat{\beta}_2, s^2)$. Finally, if σ^2 is not known and estimated by s_i^2 (the estimator of σ^2 in model $\mathcal{M}_i$), then it is no longer true that all t- and F-statistics depend only on $(\hat{\beta}_2, s^2)$. However, they still depend only on $M_1 y$, because we have seen that both $\hat{\beta}_{2i}$ and the residuals e_i from model $\mathcal{M}_i$ are linear functions of $M_1 y$. We conclude that the regularity conditions on λ_i are reasonable and mild.

The equivalence theorem proved in Danilov and Magnus (2004, Theorem 1), generalizing an earlier result in Magnus and Durbin (1999), states that if Assumption 1 holds and the regularity conditions (19) on λ_i are satisfied, then

$$\mathrm{E}(b_1) = \beta_1 - Q\,\mathrm{E}(W\hat{\beta}_2 - \beta_2),$$
$$\mathrm{var}(b_1) = \sigma^2 (X_1'X_1)^{-1} + Q\,\mathrm{var}(W\hat{\beta}_2)Q',$$

and hence

$$\mathrm{MSE}(b_1) = \sigma^2 (X_1'X_1)^{-1} + Q\,\mathrm{MSE}(W\hat{\beta}_2)Q'. \tag{20}$$

The importance of the equivalence theorem lies in the fact that the properties of the complicated WALS estimator b_1 of β_1 depend critically on the properties of the less complicated estimator $W\hat{\beta}_2$ of β_2. We notice that neither the bias, nor the variance or the mean squared error of b_1 depend on β_1. They do, however, depend on β_2.

It follows from the equivalence theorem (from (20) in particular) that the WALS estimator b_1 will be a 'good' estimator of β_1 (in the mean squared error sense) if and only if $W\hat{\beta}_2$ is a 'good' estimator of β_2. Now, under Assumption 1, the matrix W is diagonal, say $W = \mathrm{diag}(w_1, \ldots, w_{k_2})$. Suppose that σ^2 is known (we discuss the unknown σ^2 case later), and that we choose $w_j = w_j(\hat{\beta}_{2j})$. Then, since the $\{\hat{\beta}_{2j}\}$ are independent, so are the $\{w_j\hat{\beta}_{2j}\}$, and our k_2-dimensional problem reduces to k_2 (identical) one-dimensional problems: only using the information that $\hat{\beta}_{2j} \sim \mathrm{N}(\beta_{2j}, \sigma^2)$ and assuming that σ^2 is known, find the best (in the mean squared error sense) estimator of β_{2j}. The Laplace estimator discussed below solves this problem.

Suppose $\tilde{\beta}_{2j}$ is the desired optimal estimator of β_{2j}. Then, letting $\tilde{\beta}_2 := (\tilde{\beta}_{21}, \ldots, \tilde{\beta}_{2k_2})'$, the equivalence theorem directly gives us the optimal WALS estimator

$$b_1 = \hat{\beta}_{1r} - Q\tilde{\beta}_2,$$

with

$$\mathrm{E}(b_1) = \beta_1 - Q\,\mathrm{E}(\tilde{\beta}_2 - \beta_2),$$
$$\mathrm{var}(b_1) = \sigma^2 (X_1'X_1)^{-1} + Q\,\mathrm{var}(\tilde{\beta}_2)Q'.$$

From a computational point of view, it is important to note that the number of required calculations is of order k_2, even though there are 2^{k_2} models to consider. This is so because we do not need all 2^{k_2} individual λ's; only k_2 linear combinations are required, namely the diagonal elements of W; see Leamer (1978, p. 154) for a related result in terms of principal components.

3.4. The Laplace estimator

Thus motivated, let x be a single observation from a univariate normal distribution with mean η and variance one, that is, $x \sim \mathrm{N}(\eta, 1)$. How to estimate η? This seemingly trivial question was addressed in Magnus (2002). We consider five candidates (there are more):

- the 'usual' estimator: $t(x) = x$
- the 'silly' estimator: $t(x) = 0$
- the pretest estimator:

$$t(x) = \begin{cases} 0 & \text{if } |x| \le c \\ x & \text{if } |x| > c \end{cases}$$

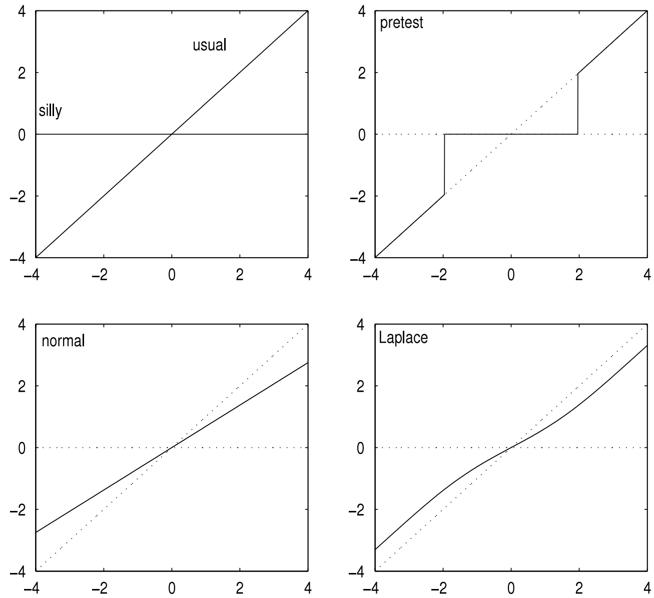

Fig. 1. Five estimators $t(x)$ of η when $x \sim \mathrm{N}(\eta, 1)$.

- the 'normal' estimator: $t(x) = x/(1 + c)$
- the Laplace estimator defined below in (21),

where c is a (generic) nonnegative constant. The five estimators are graphed in Fig. 1, where $c = 1.96$ for the pretest estimator, $c = 1/2.1981$ for the 'normal' estimator, and $c = \log(2)$ for the Laplace estimator.

The usual estimator is unbiased, admissible, and minimax. Its risk $R(\eta) := \mathrm{E}(t(x) - \eta)^2 = 1$ has good properties when $|\eta|$ is large, but not when η is close to zero. The silly estimator has excellent properties when η is close to zero, but its risk $R(\eta) = \eta^2$ increases without bound when $|\eta|$ becomes large. The pretest estimator has bounded risk, but it has a discontinuity and is therefore inadmissible. Also, its risk is higher than either the usual or the silly estimator when $|\eta|$ is around one. The 'normal' estimator is a Bayesian estimator, combining the likelihood $x|\eta \sim \mathrm{N}(\eta, 1)$ with a normal prior $\pi(\eta) \sim \mathrm{N}(0, 1/c)$. (In Fig. 1 we take $1/c = 2.1981$, so that $\Pr(|\eta| < 1) = 1/2$.) This is — in essence — the BMA estimator. The risk of the 'normal' estimator is also unbounded. The Laplace estimator was developed as an estimator which is admissible, has bounded risk, has good properties around $|\eta| = 1$, and is near-optimal in terms of minimax regret. It is a Bayesian estimator, based on the Laplace prior

$$\pi(\eta) = \frac{c}{2}\exp(-c|\eta|).$$

The hyperparameter c is chosen $c = \log 2$, because this implies that the prior median of η is zero and the prior median of η^2 is one, which comes closest to our prior idea of ignorance as discussed in the Introduction.

The moments of the posterior distribution of $\eta|x$ are given in Theorem 1, which extends Pericchi and Smith (1992) and Magnus (2002).

Theorem 1. *Consider the likelihood and prior*

$$x|\eta \sim \mathrm{N}(\eta, 1), \qquad \pi(\eta) = \frac{c}{2}\exp(-c|\eta|),$$

where c is a positive hyperparameter. Let $Q(x, \eta) := (x-\eta)^2 + 2c|\eta|$. Then the posterior distribution of η given x is given by

$$p(\eta \mid x) = \frac{\exp(-Q(x, \eta)/2)}{\int \exp(-Q(x, \eta)/2)\,\mathrm{d}\eta}.$$

The mean and variance of the posterior distribution are given by

$$\mathrm{E}(\eta \mid x) = \frac{1+h(x)}{2}(x-c) + \frac{1-h(x)}{2}(x+c) \tag{21}$$

and

$$\mathrm{var}(\eta \mid x) = 1 + c^2(1-h^2(x)) - \frac{c(1+h(x))\phi(x-c)}{\Phi(x-c)},$$

where

$$h(x) := \frac{\mathrm{e}^{-cx}\Phi(x-c) - \mathrm{e}^{cx}\Phi(-x-c)}{\mathrm{e}^{-cx}\Phi(x-c) + \mathrm{e}^{cx}\Phi(-x-c)},$$

and $\phi(x)$ and $\Phi(x)$ denote the density and cumulative distribution function of the standard-normal distribution, respectively.

Proof. Writing

$$Q(x,\eta) = \begin{cases} (\eta-(x+c))^2 - 2cx - c^2 & \text{if } \eta \le 0, \\ (\eta-(x-c))^2 + 2cx - c^2 & \text{if } \eta > 0, \end{cases}$$

and realizing that

$$\int_{-\infty}^{x} t\phi(t)\,\mathrm{d}t = -\phi(x), \qquad \int_{-\infty}^{x} t^2\phi(t)\,\mathrm{d}t = \Phi(x) - x\phi(x),$$

the posterior distribution follows easily.

The moments are easy to compute. Note that the function h is monotonically increasing with $h(-\infty) = -1$, $h(0) = 0$, and $h(\infty) = 1$, and that $h(-x) = -h(x)$. □

3.5. Implementation using Laplace priors

The WALS estimation procedure can be summarized as follows.

- In the unrestricted model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, determine which are the focus regressors X_1 and which are the auxiliary regressors X_2.
- Compute $M_1 := I_n - X_1(X_1'X_1)^{-1}X_1'$, and then P (orthogonal) and Λ (diagonal) such that $P'X_2'M_1X_2P = \Lambda$. Compute $X_2^* := X_2P\Lambda^{-1/2}$, so that $X_2^{*\prime}M_1X_2^* = I_{k_2}$. Letting $\beta_2^* := \Lambda^{1/2}P'\beta_2$, note that $X_2^*\beta_2^* = X_2\beta_2$.
- Compute $\hat{\beta}_2^{\,*} = X_2^{*\prime}M_1y$.
- Let $\eta := \beta_2^*/\sigma$. Assuming that σ^2 is known, compute $\hat{\eta} := \hat{\beta}_2^{\,*}/\sigma$. Notice that the components $\hat{\eta}_1, \ldots, \hat{\eta}_{k_2}$ of $\hat{\eta}$ are independent and that $\hat{\eta}_j \sim \mathrm{N}(\eta_j, 1)$.
- For $j = 1, \ldots, k_2$ compute the Laplace estimator $\tilde{\eta}_j := \mathrm{E}(\eta_j|\hat{\eta}_j)$ and its variance $\omega_j^2 := \mathrm{var}(\eta_j|\hat{\eta}_j)$. Define $\tilde{\eta} := (\tilde{\eta}_1, \ldots, \tilde{\eta}_{k_2})'$ and $\Omega := \mathrm{diag}(\omega_1^2, \ldots, \omega_{k_2}^2)$.
- Since $\eta = \beta_2^*/\sigma = \Lambda^{1/2}P'\beta_2/\sigma$, we obtain $\beta_2 = \sigma P\Lambda^{-1/2}\eta$, and hence we compute the WALS estimators for β_2 and β_1 as

  $$b_2 = \sigma P\Lambda^{-1/2}\tilde{\eta}, \qquad b_1 = (X_1'X_1)^{-1}X_1'(y - X_2b_2).$$

- Letting $Q := (X_1'X_1)^{-1}X_1'X_2$, the variance of b_2 and b_1 is

  $$\mathrm{var}(b_2) = \sigma^2P\Lambda^{-1/2}\Omega\Lambda^{-1/2}P',$$

  and

  $$\mathrm{var}(b_1) = \sigma^2(X_1'X_1)^{-1} + Q\,\mathrm{var}(b_2)Q'.$$

  We also have $\mathrm{cov}(b_1, b_2) = -Q\,\mathrm{var}(b_2)$. In standard applications one is primarily interested in the diagonal elements of the variance matrices.

Finally, we note that we have assumed that σ^2 is known, whereas in fact it is of course not known. Our solution to this problem is to replace σ^2 by s^2, the estimate in the unrestricted model. This is an approximation, but a very accurate one, as demonstrated and exemplified by Danilov (2005).

3.6. BMA and WALS compared

It may seem at first glance that the two estimation procedures BMA and WALS are quite different, but in fact they are conceptually quite close. Both procedures are model averaging algorithms. The assumption that the data are normally distributed is the same, and the treatment of the focus parameters β_1 and the error variance σ^2 as noninformative priors is essentially the same. The difference between BMA and WALS lies in the prior treatment of the auxiliary parameters β_2. In BMA we assume normality of the priors with

$$\mathrm{E}(\beta_{2i} \mid \mathcal{M}_i) = 0, \qquad \mathrm{var}(\beta_{2i} \mid \mathcal{M}_i) = \frac{\sigma^2}{g}(X_{2i}'M_1X_{2i})^{-1},$$

where $g := 1/\max(n, k_2^2)$. Since $\beta_{2i} = T_i'\beta_2$, $X_{2i} = X_2T_i$, and $T_i'T_i = I_{k_{2i}}$, we can write these moments as

$$\mathrm{E}(\beta_2 \mid \mathcal{M}_i) = 0, \qquad \mathrm{var}(\beta_2 \mid \mathcal{M}_i) = \frac{\sigma^2}{g}T_i(T_i'X_2'M_1X_2T_i)^{-1}T_i'. \tag{22}$$

In contrast, in WALS we write β_2 in terms of η as $\beta_2 = \sigma P\Lambda^{-1/2}\eta$. The k_2 components of η are i.i.d. according to a Laplace distribution

$$\pi(\eta_i) = \frac{c}{2}\exp(-c|\eta_i|), \quad c = \log 2.$$

This implies that each η_i is symmetrically distributed around zero, that the median of η_i^2 is one, and that the variance of η_i is $\sigma_\eta^2 = 2/c^2$. This choice of prior moments is based on our idea of ignorance as a situation where we do not know whether the theoretical t-ratio is larger or smaller than one in absolute value. The prior moments of β_2 are then given by

$$\mathrm{E}(\beta_2) = 0, \qquad \mathrm{var}(\beta_2) = \sigma^2\sigma_\eta^2P\Lambda^{-1}P' = \frac{\sigma^2}{c^2/2}(X_2'M_1X_2)^{-1}. \tag{23}$$

Comparing (22) and (23) shows that these prior moments are in fact closely related, and suggests in addition a new value for g in BMA applications, namely $g = c^2/2 = 0.24$.

The conceptual differences are thus the distribution (Laplace versus normal), where Laplace has the advantage of leading to finite risk; and the choice of g as a scaling parameter for the prior variance.

4. Growth models

In the neoclassical growth model (Solow, 1956), growth around a steady state is determined by rates of physical capital accumulation, population growth, and exogenous technological progress. The initial income of an economy is relevant for its transition path as countries with a lower initial income are expected to grow faster than richer countries. The 'new growth' theories seek to explain also the previously exogenous components of economic growth, which is why they are often called 'endogenous' growth models. A frequently used empirical model for growth regressions is the human capital-augmented neoclassical model (Mankiw et al., 1992), which regresses the average growth rate of GDP per capita on investment, the log of initial GDP per capita, the population growth rate, and a human capital variable.

The 'Solow' determinants derived from a neoclassical growth model are sometimes called 'proximate' determinants because they are thought to be the most established drivers of economic growth. The term 'proximate' also reflects the ease with which these determinants can be influenced by policy measures, thus emphasizing their importance for empirical research and policy advice. Recent literature advocates the view that these proximate determinants in turn depend on slow-moving 'fundamental' growth determinants such as a country's geography, the quality of its institutions, the degree of fractionalization in its society, and its culture or religion; see Durlauf et al. (2008a) and references therein. Hall and Jones (1999) use a similar framework in which they distinguish between proximate causes of economic success (capital accumulation and productivity) and a more fundamental determinant which they name 'social structure'. Accordingly, one can distinguish between proximate and fundamental growth *theories* (Durlauf et al., 2008b).

We seek to capture these different types of growth determinants (theories) in our empirical analysis. Thus we define two different sets of regressors, labeled X_1 and X_2, somewhat in the spirit of Brock and Durlauf (2001) and Masanjala and Papageorgiou (2008). The set X_1 contains the regressors which appear in every regression on theoretical or other grounds (irrespective of their statistical significance), the so-called 'focus' regressors. Typically, but not necessarily, X_1 contains the constant term as one of its regressors. The additional controls in the regression, the so-called 'auxiliary' regressors, are contained in X_2. Their primary role is to improve the estimation of the focus regressors, although their estimates may be of independent interest. The distinction between focus and auxiliary regressors is helpful when one wants to understand the relationship between neoclassical and other new growth determinants. While the Solow variables appear in many empirical studies, thus serving as a baseline for growth analysis, it is not so clear which variables should be included as auxiliary regressors. The proximate (Solow) determinants are the variables of major interest in our analysis (X_1) based on their prominent position in growth theory and growth empirics. The fundamental growth determinants mentioned above are included as an additional set of regressors (X_2) serving as controls of the standard growth models.[4]

We analyze two different model specifications: Model 1 and Model 2. In the notion of Durlauf et al. (2008b), we interpret Model 1 as a direct test of the proximate neoclassical growth theory against the fundamental new growth theories of institutions, geography, fractionalization, and religion. Model 2 deviates from the proximate versus fundamental classification, and tests the robustness of the endogenous growth model using the distinction between focus and auxiliary regressors.

In both models the dependent variable is GROWTH. In our data set, the average growth rate is 1.99% with a standard error 1.86. The regressors and their role as either focus or auxiliary are given in Table 1.

Model 1 contains six focus regressors (including the constant term) and four auxiliary regressors. It is motivated by the neoclassical growth model and thus contains all Solow determinants as focus regressors (X_1). These are: The initial capital stock of an economy (GDP60), measured as the log of GDP per capita in 1960. This represents the so-called convergence term of the Solow growth model and attempts to analyze whether poorer countries (those having lower initial income) actually grow faster than richer ones. Next, the 1960–1985 equipment investment share of GDP (EQUIPINV), which serves as a proxy for the stock of physical capital in the economy and reflects the importance of capital accumulation for the growth of an economy. Then two variables which represent human capital. To capture different facets of human capital, we include a direct measure, the total gross enrollment rate in primary schooling in 1960 (SCHOOL60), and also a proxy for noneducational human capital, the life expectancy at age zero, measured in 1960 (LIFE60). Both human capital variables are widely used proxies for the initial human capital stock in an economy and are expected to have a positive effect on productivity growth with life expectancy being the more robust regressor (Sala-i-Martin, 1997). Whenever possible, we use such initial values for our variables, reducing also the potential endogeneity problem in our growth regressions. Finally, the population growth rate between 1960 and 1990 (DPOP), a proxy for the exogenous growth rate of labor assumed to foster productivity growth in the neoclassical model.

Table 1
Model specifications, focus and auxiliary regressors.

Variable	Model 1	Model 2	Mean	SE
CONSTANT	Focus	Focus	1.0000	0.0000
GDP60	Focus	Focus	7.5253	0.8612
EQUIPINV	Focus	Focus	0.0432	0.0344
SCHOOL60	Focus	Focus	0.7807	0.2556
LIFE60	Focus	Focus	56.0676	1.1566
DPOP	Focus	Auxiliary	0.0206	0.0100
LAW	Auxiliary	Focus	0.5518	0.3332
TROPICS	Auxiliary	Focus	0.5481	0.4709
AVELF	Auxiliary	Focus	0.2984	0.2797
CONFUC	Auxiliary	Focus	0.0185	0.0862
MINING	–	Auxiliary	0.0482	0.0792
PRIGHTS	–	Auxiliary	3.4551	1.9073
MALARIA	–	Auxiliary	0.2866	0.4036

To test this neoclassical model (theory) and its proximate growth determinants we include the suggested fundamental growth determinants as auxiliary regressors. There is not only theoretical but also empirical support for these regressors; see Sala-i-Martin (1997), Fernández et al. (2001a) and Sala-i-Martin et al. (2004). We specify the following set of four auxiliary variables in X_2. First, a rule of law index (LAW), a measure of the importance of institutions, supposed to have a positive effect on economic growth. Next, a country's fraction of tropical area (TROPICS), which controls for the effect of geography and is expected to have a negative effect on productivity growth. Third, an average index of ethnolinguistic fragmentation in a country (AVELF), which will help to analyze the influence of the degree of fractionalization in society and culture on economic productivity, typically found to be negative. And finally, the fraction of Confucian population in a country (CONFUC), used as a (somewhat dubious) proxy for culture or religion, typically identified as having a positive effect on growth. CONFUC can also be viewed as a proxy for the 'Asian (baby) tigers': Hong Kong, Malaysia, Singapore, South Korea, and Taiwan. However, as not all of the Asian countries with large growth rates are Confucian, CONFUC is more than just a regional dummy.

Model 2 contains nine focus regressors and four auxiliary regressors, and it represents an endogenous growth model trying to identify more specifically the factors driving growth and technological progress than is possible in Model 1. All regressors of our first model are included in Model 2 as focus regressors, except DPOP which is now an auxiliary regressor, because of its ambiguous role in economic growth. This ambiguity and lack of robustness was found, for example, by Sala-i-Martin (1997), Fernández et al. (2001a), and Sala-i-Martin et al. (2004). Our results reported in Section 5 confirm this ambiguity.

The three new auxiliary regressors are: the fraction of GDP produced in mining (MINING), a structural variable supposed to exert a negative effect on economic growth; an index for political rights (PRIGHTS), serving as a second institutional variable (the other is LAW), so that we capture not only the quality of the legal framework in a country but also a notion of public participation in the political process; and malaria prevalence in 1966 (MALARIA), another geographical variable (next to TROPICS), so that we account not only for the geographical location of a country, but also for its disease environment.

[4] Our set-up is distantly related to the empirical study by Levine and Renelt (1992) who include (as we do) a set of variables that appear in every regression. They distinguish, however, between *three* sets of variables with the aim of finding the widest range of coefficient estimates on the variables of interest that standard hypothesis tests do not reject, thus assessing the robustness of partial correlations between the per capita growth rate and various economic indicators.

Table 2
Estimates $\hat{\beta}$ and standard errors (in parentheses), Model 1, Set-up 1.

Regressor	Unrestricted	Restricted	GtS	WALS	BMA
Focus regressors					
CONSTANT	0.0609 (0.0223)	0.0587 (0.0242)	0.0518 (0.0214)	0.0594 (0.0221)	0.0492 (0.0229)
GDP60	−0.0155 (0.0033)	−0.0160 (0.0035)	−0.0145 (0.0032)	−0.0156 (0.0033)	−0.0139 (0.0035)
EQUIPINV	0.1366 (0.0552)	0.2405 (0.0583)	0.1377 (0.0555)	0.1555 (0.0551)	0.1644 (0.0615)
SCHOOL60	0.0170 (0.0098)	0.0184 (0.0111)	0.0191 (0.0097)	0.0175 (0.0097)	0.0160 (0.0102)
LIFE60	0.0008 (0.0004)	0.0010 (0.0004)	0.0008 (0.0004)	0.0009 (0.0004)	0.0008 (0.0004)
DPOP	0.3466 (0.2503)	−0.0341 (0.2611)	0.3275 (0.2513)	0.2651 (0.2487)	0.1654 (0.2770)
Auxiliary regressors					
LAW	0.0174 (0.0066)	–	0.0167 (0.0066)	0.0147 (0.0065)	0.0109 (0.0093)
TROPICS	−0.0075 (0.0040)	–	−0.0083 (0.0039)	−0.0055 (0.0037)	−0.0035 (0.0047)
AVELF	−0.0077 (0.0058)	–	–	−0.0053 (0.0048)	−0.0021 (0.0047)
CONFUC	0.0562 (0.0164)	–	0.0596 (0.0163)	0.0443 (0.0163)	0.0612 (0.0185)

Table 3
Estimates $\hat{\beta}$ and standard errors (in parentheses), Model 1, Set-up 2.

Regressor	Unrestricted	Restricted	GtS	WALS	BMA
Focus regressor					
CONSTANT	0.0609 (0.0223)	0.0199 (0.0022)	0.0344 (0.0146)	0.0560 (0.0215)	0.0488 (0.0218)
Auxiliary regressors					
GDP60	−0.0155 (0.0033)	–	−0.0120 (0.0032)	−0.0136 (0.0033)	−0.0129 (0.0040)
EQUIPINV	0.1366 (0.0552)	–	0.1951 (0.0524)	0.1037 (0.0537)	0.1539 (0.0797)
SCHOOL60	0.0170 (0.0098)	–	–	0.0125 (0.0094)	0.0084 (0.0127)
LIFE60	0.0008 (0.0004)	–	0.0012 (0.0003)	0.0008 (0.0003)	0.0009 (0.0005)
DPOP	0.3466 (0.2503)	–	–	0.2236 (0.2156)	0.0261 (0.1252)
LAW	0.0174 (0.0066)	–	–	0.0137 (0.0063)	0.0090 (0.0092)
TROPICS	−0.0075 (0.0040)	–	–	−0.0055 (0.0039)	−0.0021 (0.0038)
AVELF	−0.0077 (0.0058)	–	–	−0.0083 (0.0057)	−0.0024 (0.0050)
CONFUC	0.0562 (0.0164)	–	0.0728 (0.0167)	0.0451 (0.0163)	0.0663 (0.0180)

Table 4
Estimates $\hat{\beta}$ and standard errors (in parentheses), Model 2, Set-up 1.

Regressor	Unrestricted	Restricted	GtS	WALS	BMA
Focus regressors					
CONSTANT	0.0931 (0.0264)	0.0768 (0.0193)	0.0930 (0.0198)	0.0879 (0.0246)	0.0862 (0.0239)
GDP60	−0.0173 (0.0033)	−0.0156 (0.0033)	−0.0166 (0.0032)	−0.0167 (0.0033)	−0.0164 (0.0033)
EQUIPINV	0.1324 (0.0579)	0.1479 (0.0550)	0.1448 (0.0531)	0.1379 (0.0562)	0.1423 (0.0553)
CONFUC	0.0538 (0.0169)	0.0585 (0.0165)	0.0522 (0.0161)	0.0550 (0.0167)	0.0550 (0.0169)
SCHOOL60	0.0144 (0.0096)	0.0183 (0.0098)	0.0151 (0.0096)	0.0156 (0.0096)	0.0162 (0.0099)
LIFE60	0.0006 (0.0004)	0.0006 (0.0003)	0.0005 (0.0003)	0.0006 (0.0003)	0.0006 (0.0003)
LAW	0.0200 (0.0068)	0.0145 (0.0063)	0.0183 (0.0063)	0.0183 (0.0066)	0.0171 (0.0067)
TROPICS	−0.0055 (0.0041)	−0.0055 (0.0037)	−0.0029 (0.0037)	−0.0053 (0.0040)	−0.0044 (0.0041)
AVELF	−0.0040 (0.0060)	−0.0073 (0.0059)	−0.0033 (0.0059)	−0.0049 (0.0059)	−0.0050 (0.0062)
Auxiliary regressors					
MINING	−0.0090 (0.0192)	–	–	−0.0056 (0.0149)	−0.0003 (0.0063)
DPOP	0.3352 (0.2542)	–	–	0.2147 (0.2178)	0.0650 (0.1705)
PRIGHTS	−0.0013 (0.0012)	–	–	−0.0008 (0.0010)	−0.0002 (0.0007)
MALARIA	−0.0104 (0.0052)	–	−0.0122 (0.0051)	−0.0075 (0.0050)	−0.0072 (0.0070)

Table 5
Estimates $\hat{\beta}$ and standard errors (in parentheses), Model 2, Set-up 2.

Regressor	Unrestricted	Restricted	GtS	WALS	BMA
Focus regressor					
CONSTANT	0.0931 (0.0264)	0.0199 (0.0022)	0.0828 (0.0183)	0.0897 (0.0252)	0.0796 (0.0251)
Auxiliary regressors					
GDP60	−0.0173 (0.0033)	–	−0.0163 (0.0032)	−0.0156 (0.0033)	−0.0151 (0.0039)
EQUIPINV	0.1324 (0.0579)	–	0.1414 (0.0530)	0.0962 (0.0552)	0.1341 (0.0805)
CONFUC	0.0538 (0.0169)	–	0.0558 (0.0160)	0.0397 (0.0161)	0.0562 (0.0197)
SCHOOL60	0.0144 (0.0096)	–	–	0.0098 (0.0088)	0.0079 (0.0115)
LIFE60	0.0006 (0.0004)	–	0.0008 (0.0003)	0.0006 (0.0003)	0.0006 (0.0005)
LAW	0.0200 (0.0068)	–	0.0177 (0.0062)	0.0177 (0.0065)	0.0169 (0.0095)
TROPICS	−0.0055 (0.0041)	–	–	−0.0048 (0.0039)	−0.0006 (0.0021)
AVELF	−0.0040 (0.0060)	–	–	−0.0054 (0.0058)	−0.0006 (0.0027)
MINING	−0.0090 (0.0192)	–	–	−0.0063 (0.0182)	−0.0002 (0.0062)
DPOP	0.3352 (0.2542)	–	–	0.2146 (0.2175)	0.0179 (0.0990)
PRIGHTS	−0.0013 (0.0012)	–	–	−0.0012 (0.0010)	−0.0002 (0.0006)
MALARIA	−0.0104 (0.0052)	–	−0.0151 (0.0046)	−0.0090 (0.0048)	−0.0129 (0.0068)