JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12094

WEIGHTED-AVERAGE LEAST SQUARES (WALS): A SURVEY

Jan R. Magnus*
VU University Amsterdam

Giuseppe De Luca
University of Palermo

Abstract. Model averaging has become a popular method of estimation, following increasing evidence that model selection and estimation should be treated as one joint procedure. Weighted-average least squares (WALS) is a recent model-average approach, which takes an intermediate position between frequentist and Bayesian methods, allows a credible treatment of ignorance, and is extremely fast to compute. We review the theory of WALS and discuss extensions and applications.

Keywords. Computing time; Frequentist versus Bayesian; Least squares; Model averaging; Priors

1. Introduction

Our story begins with the t-ratio. Let us consider the model

$$y_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \epsilon_j \qquad (j = 1, \ldots, n)$$

We obtain estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, and their estimated variances $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. Next, we consider the t-ratio $t_2 = \hat{\beta}_2/\hat{\sigma}_2$. This t-ratio can be viewed in two ways. We could be interested in testing the hypothesis that $\beta_2 = 0$. In that case t_2 can be fruitfully employed, because under certain assumptions (such as normality) the t-ratio follows a Student distribution under the null hypothesis and if we fix the significance level of the test (say at 5%) then we can reject or not reject the hypothesis.

But the t-ratio is also commonly employed in a different way. Suppose we are primarily interested in the value of β_1. Then t_2 is often used as a diagnostic, rather than as a test statistic, in order to decide whether we wish to keep x_2 in the model or not. In this situation the 5% level is also typically used, although we could equally well argue in favor of the 95% or the 50% level or any other percentage. The two situations are different because in the first case we are interested in β_2 while in the second case we are interested in β_1. In the first case we ask: Is it true that $\beta_2 = 0$? In the second case: Does inclusion of x_2 improve the estimator of β_1? Two different questions requiring two different approaches.

*Corresponding author contact email: jan@janmagnus.nl; Tel: +31-20-598-6010.

Journal of Economic Surveys (2016) Vol. 30, No. 1, pp. 117–148

Let us consider the second situation, where t_2 is used as a diagnostic, in more detail. We have three estimators of β_1, namely the estimator from the unrestricted model, $\hat{\beta}_{1u}$; the estimator from the restricted model (where $\beta_2 = 0$), $\hat{\beta}_{1r}$; and the estimator after the preliminary test,

$$b_1 = w\hat{\beta}_{1u} + (1-w)\hat{\beta}_{1r}, \qquad w = \begin{cases} 1 & \text{if } |t_2| > c, \\ 0 & \text{if } |t_2| \leq c, \end{cases}$$

for some $c > 0$, such as $c = 1.96$ corresponding to the 5% level. This is the pretest estimator. The pretest estimator is "kinked," which has both theoretical and practical consequences (Judge and Bock, 1978; Giles and Giles, 1993). A theoretical drawback is that the estimator is inadmissible, because any estimator which is not differentiable (worse still, discontinuous) is inadmissible (Magnus, 1999). A related practical problem—familiar to all empirical economists—is the property that for $t_2 = 1.95$ we choose one estimator and for $t_2 = 1.97$ another, while in fact there is little difference between 1.95 and 1.97. These considerations lead us to reconsider the estimator $b_1 = w\hat{\beta}_{1u} + (1-w)\hat{\beta}_{1r}$ by allowing w to be a smoothly increasing function of $|t_2|$. This is model averaging in its simplest form, and we see that it is just the continuous counterpart to pretesting.

To bring out the difference between pretesting and model averaging, suppose a king has 12 advisors. He wishes to forecast next year's inflation and calls each of the advisors in for consultation. He knows his advisors and obviously has more faith in some than in others. All 12 deliver their forecast, and the king is left with 12 numbers. How to choose from these 12 numbers? Let us consider two possibilities (there are more). The king could argue: which advisor do I trust most, whom do I believe is most competent? Then I take his or her advice. The king could also argue: all advisors have something useful to say, although not in the same degree. Some are cleverer and more informed than others and their forecast should get a higher weight. Which way of thinking is better? Intuitively most people prefer the second method (model averaging), where all pieces of advice are taken into account. In standard econometrics, however, it is the first method (pretesting) which dominates.

In practice, econometricians use not one or two, but many models. If we use diagnostic tests to search for the best-fitting model, then we need to take into account not only the uncertainty of the estimates in the selected model, but also the fact that we have used the data to select a model. In other words, model selection and estimation should be seen as a combined effort, not as two separate efforts, and failure to do so may lead to misleadingly precise estimates.

In pretesting one typically reports the properties of the estimator as if estimation had not been preceded by model selection. Standard statistical theory is therefore not directly applicable, since the properties of pretest estimators depend not only on the stochastic nature of the selected model, but also on the way the model has been selected. Problems associated with inference after model selection have been investigated in Magnus (1999, 2002), Danilov and Magnus (2004a, b), Leeb and Pötscher (2003, 2005, 2006, 2008), and others. All these studies conclude that ignoring the uncertainty associated with the model selection step can lead to seriously misleading inference. Another major drawback of ignoring the noise produced by model selection is that small perturbations of the data may result in very different models being selected (Yang, 2001).

In model averaging one does not select a single model out of the available set of models—in fact, the question "which is the correct model" is not answered, because model selection is thought of as an intermediate step, not a goal in itself. Each model contributes information on the parameters of interest, and all these pieces of information are combined into an unconditional estimate using a weighted average of the conditional estimates across all possible models. The theory of model averaging thus incorporates the uncertainty arising from estimation and model selection jointly. (Model averaging is not the only method which combines estimation uncertainty and model selection uncertainty—penalized regression is also widely used (see the discussion in Kumar and Magnus (2013, p. 221))).

One can estimate the parameters from either a frequentist or a Bayesian perspective. Also, one can choose the weights from a frequentist or a Bayesian perspective. This gives rise to four different types

of model averaging. The method presented here, called weighted-average least squares (WALS), is a Bayesian combination of frequentist estimators, and it has advantages over other model-average methods that will be discussed later.

The theory and application of WALS was developed in a large number of papers. In many of these papers we did not immediately find the shortest proofs (this is true in particular for the equivalence theorem) or the most suitable prior (Laplace, Weibull), there are some ambiguities and misunderstandings that need to be put right (for example about the semiorthogonal transformation, originally introduced as an assumption—which it is not), and the notation is not consistent across papers. In this review, we attempt to present one consistent theory of WALS. This review paper is, however, rather more than a summary of past results. For example, much attention is given to the underlying assumptions and where and why precisely they are needed, and a new improved prior is introduced for the weight function.

The paper is organized as follows. In Sections 2 and 3, we present the framework and the constrained least-squares estimators. WALS and the equivalence theorem are introduced in Section 4. After an interlude on the question whether weights should necessarily lie between zero and one (Section 5), we discuss preliminary scaling, the semiorthogonal transformation, and its consequences (Sections 6 and 7). Then we turn to the weights by providing a frequentist weight function in Section 8 and a Bayesian weight function in Section 9. We prefer the latter because it allows us to obtain an admissible estimator which also has a credible interpretation in terms of ignorance. The analysis so far has assumed that the error variance in the regression model is known. This is clearly unrealistic and Section 10 discusses how to deal with this additional problem. Section 11 puts it all together and provides a 7-step outline of the procedure.

Next we discuss extensions (Section 12), some of which have already been analyzed in the WALS literature, while some are suggestions for future research. We also highlight some empirical applications of the WALS method. In Section 13, we compare WALS with other model-average estimators and in Section 14 we discuss user-friendly software for WALS, both in `MATLAB` and `Stata`.

2. Framework and Preliminaries

Our data are assumed to be generated by the linear process

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{1}$$

where y $(n \times 1)$ is the vector of observations on the outcome of interest, 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 nonrandom parameter vectors. We assume that $k_1 \geq 1$, $k_2 \geq 1$, $k = k_1 + k_2 \leq n - 1$, that $X = (X_1 : X_2)$ has full column-rank k, and that the disturbances $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent and identically distributed (i.i.d.) and normal, so that $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$.

This is the classical setup of the linear model (including fixed regressors and normality), except that we distinguish between two types of regressors: X_1 and X_2. The reason for distinguishing between X_1 and X_2 is that X_1 contains explanatory variables which we want to be present in each model on theoretical or other grounds (irrespective of the observed t-values of the β_1-parameters), while X_2 contains additional explanatory variables of which we are less certain. This setup is more general than the conventional case where typically all regressors are auxiliary except the constant term. The new setup thus allows the investigator to keep a regressor in the model even when diagnostic tests suggest to remove it. The columns of X_1 are called "focus" regressors and the columns of X_2 "auxiliary" regressors. Similarly, the components of β_1 are called "focus" parameters, and the components of β_2 "auxiliary" parameters.

In the context of estimation (we shall also consider prediction), our interest is in the estimation of β_1, and the only role for X_2 is to "improve" the estimator of β_1. However, if the investigator is interested in both β_1 *and* β_2, then the analysis and the model-average properties do not change. This is the key

message of Proposition 4.2 (the equivalence theorem), and we shall expand on this surprising result in Section 4.

The data-generation process (DGP) described in (1) is of course not known to the investigator, and hence the model or models used by him or her will in general deviate from it. We shall only consider models that are *smaller* than (or equal to) the DGP, so that the model space $\mathcal{M}$ consists of all submodels $\mathcal{M}_i$ of (1) that contain *all* focus regressors and *some* of the auxiliary regressors. In practice, the DGP and the largest model will not coincide. There may be a set of regressors, say X_3, such that the DGP is given by

$$y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \epsilon,$$

where X_1 is always in the model, X_2 may or may not be in the model and X_3 is never in the model. This is more general and more realistic than (1). The regressors in X_3 may not be known to us, or we may know them in theory but lack the data, or we may know them in theory and have the relevant data but believe *a priori* that the associated β_3-parameter is so "small" that the increase in misspecification bias is outweighed by the decrease in estimated standard errors. In the first two cases we cannot include X_3 in the DGP, but in the third case we can and we should so that X_3 would belong to the DGP but not to the model space. In what follows we shall not follow this route and make the simplifying assumption that (1) is the DGP, so that the largest model and the DGP coincide. Model uncertainty is thus restricted to a well-defined class of models which is known in advance, the so-called $\mathcal{M}$-closed perspective (Hoeting *et al.*, 1999).

When $k_2 = 1$ (one auxiliary regressor), we have two models: the unrestricted and the restricted (where $\beta_2 = 0$). When $k_2 = 2$ (two auxiliary regressors), there are four possible models: the unrestricted model, two partially restricted models (one of the two components of β_2 is zero) and the fully restricted model (both components of β_2 are zero). In general, there are 2^{k_2} models to consider. The ith model $\mathcal{M}_i$ is characterized by a $k_2 \times r_i$ selection matrix S_i with rank $0 \le r_i \le k_2$, so that $S_i' = (I_{r_i} : 0)$ or a column-permutation thereof. In other words, $\mathcal{M}_i$ is defined as the linear model (1) under the restriction $S_i'\beta_2 = 0$. The number r_i denotes the number of excluded auxiliary variables and the matrix S_i specifies which r_i variables are excluded. For given r_i, there are $\binom{k_2}{r_i}$ different possible choices for the selection matrix S_i, in total

$$\sum_{r_i=0}^{k_2} \binom{k_2}{r_i} = 2^{k_2}.$$

It will be useful to define

$$M_1 = I_n - X_1(X_1'X_1)^{-1}X_1', \qquad Q = (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1/2}. \tag{2}$$

Our first interest is in the constrained least-squares (LS) estimators of β_1 and β_2 in model $\mathcal{M}_i$. As a preliminary to the general results presented in Section 3, we first consider the two extremes: the (fully) restricted model and the unrestricted model. In the restricted model (where $\beta_2 = 0$) we find

$$\hat{\beta}_{1r} = (X_1'X_1)^{-1}X_1'y, \tag{3}$$

while in the unrestricted model we have

$$\hat{\beta}_{1u} = \hat{\beta}_{1r} - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_{2u}, \qquad \hat{\beta}_{2u} = (X_2'M_1X_2)^{-1}X_2'M_1y. \tag{4}$$

The subscripts "u" and "r" denote "unrestricted" and "restricted" (with $\beta_2 = 0$), respectively. It is easy to see that $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ are always correlated, and that $\hat{\beta}_{1u}$ and $\hat{\beta}_{2u}$ are only uncorrelated when $X_1'X_2 = 0$. In contrast, $\hat{\beta}_{1r}$ and $\hat{\beta}_{2u}$ are always uncorrelated.

Proposition 2.1. *The two least-squares estimators $\hat{\beta}_{1r}$ and $\hat{\beta}_{2u}$ are independent.*

Proof. Since $M_1X_1 = 0$ we have $cov(X_1'y, X_2'M_1y) = \sigma^2 X_1'M_1X_2 = 0$, and the result follows. □

This simple fact will play an important role in the sequel. We introduce

$$\hat{\theta} = (X_2'M_1X_2)^{-1/2}X_2'M_1y, \qquad \theta = (X_2'M_1X_2)^{1/2}\beta_2, \tag{5}$$

and note that $\hat{\theta} \sim \mathrm{N}(\theta,\ \sigma^2 I_{k_2})$, so that the components of $\hat{\theta}$ are independent. The estimators in (4) now simplify to

$$\hat{\beta}_{1u} = \hat{\beta}_{1r} - Q\hat{\theta}, \qquad \hat{\beta}_{2u} = (X_2'M_1X_2)^{-1/2}\hat{\theta}, \tag{6}$$

and Proposition 2.1 tells us that $\hat{\beta}_{1r}$ and $\hat{\theta}$ are independent. These two vectors are independent because of the normality of y and the fact that $X_1'y$ and $X_2'M_1y$ are uncorrelated. In fact, even if the observations $y_1, \ldots, y_n$ are not normal and the data-generating process is unknown, $\hat{\beta}_{1r}$ and $\hat{\theta}$ will still be uncorrelated, as long as $y_1, y_2, \ldots, y_n$ are uncorrelated with constant variance (Leeb and Pötscher, 2003, lemma A.1).

3. Constrained Least Squares

Let us now consider the general case, that is, the constrained LS estimation of β_1 and β_2 in model $\mathcal{M}_i$ under the constraint $S_i'\beta_2 = 0$.

Proposition 3.1. *The LS estimators of β_1 and β_2 in model $\mathcal{M}_i$ are given by*

$$\hat{\beta}_{1(i)} = \hat{\beta}_{1r} - QW_i\hat{\theta}, \qquad \hat{\beta}_{2(i)} = (X_2'M_1X_2)^{-1/2}W_i\hat{\theta}, \tag{7}$$

where Q is given in (2)*, $W_i = I_{k_2} - P_i$, and*

$$P_i = (X_2'M_1X_2)^{-1/2}S_i\left(S_i'(X_2'M_1X_2)^{-1}S_i\right)^{-1}S_i'(X_2'M_1X_2)^{-1/2} \tag{8}$$

is a symmetric idempotent $k_2 \times k_2$ matrix of rank r_i.

Proof. This follows from Lemma A1 in Danilov and Magnus (2004a). □

In the special case where $r_i = k_2$, we have $P_i = I_{k_2}$ and $W_i = 0$: the restricted case. In the other extreme where $r_i = 0$, we have $P_i = 0$ and $W_i = I_{k_2}$: the unrestricted case. The subscript i always refers to the ith model ($i = 1, \ldots, 2^{k_2}$). We write (i) instead of i when the object is a random variable or a random vector.

We saw in (6) that the unrestricted estimator $\hat{\beta}_{1u}$ is a linear function of $\hat{\beta}_{1r}$ and $\hat{\theta}$. Proposition 3.1 shows that this remains true, more generally, for each $\hat{\beta}_{1(i)}$. The estimator $\hat{\beta}_{2(i)}$ is a linear function of $\hat{\theta}$ only and hence independent of $\hat{\beta}_{1r}$.

The distribution of the two estimators and the vector of residuals are given in the next two propositions.

Proposition 3.2. *The distribution of $\hat{\beta}_{1(i)}$ in model $\mathcal{M}_i$ is given by*

$$\hat{\beta}_{1(i)} \sim \mathrm{N}\left(\beta_1 + QP_i\theta,\ \sigma^2\left((X_1'X_1)^{-1} + QW_iQ'\right)\right),$$

the distribution of $\hat{\beta}_{2(i)}$ by

$$\hat{\beta}_{2(i)} \sim \mathrm{N}\left((X_2'M_1X_2)^{-1/2}W_i\theta,\ \sigma^2\left((X_2'M_1X_2)^{-1/2}W_i(X_2'M_1X_2)^{-1/2}\right)\right),$$

and the covariance of $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}$ is

$$cov(\hat{\beta}_{1(i)}, \hat{\beta}_{2(i)}) = -\sigma^2 QW_i(X_2'M_1X_2)^{-1/2}.$$

Proof. This follows from Proposition 3.1 and the fact that

$$\hat{\beta}_{1r} \sim \mathrm{N}(\beta_1 + Q\theta,\ \sigma^2(X_1'X_1)^{-1}), \qquad \hat{\theta} \sim \mathrm{N}(\theta,\ \sigma^2 I_{k_2}),$$

and $\hat{\beta}_{1r}$ and $\hat{\theta}$ are independent. □

Proposition 3.3. *The residual vector in model $\mathcal{M}_i$ is*

$$e_{(i)} = y - X_1\hat{\beta}_{1(i)} - X_2\hat{\beta}_{2(i)} = D_i y,$$

where

$$D_i = M_1 - M_1X_2(X_2'M_1X_2)^{-1/2}W_i(X_2'M_1X_2)^{-1/2}X_2'M_1$$

is a symmetric idempotent matrix of rank $n-k+r_i$, and the distribution of $s^2_{(i)} = e'_{(i)}e_{(i)}/(n-k+r_i)$ is given by

$$\frac{(n-k+r_i)s^2_{(i)}}{\sigma^2} \sim \chi^2\left(n-k+r_i,\ \frac{\theta' P_i\theta}{\sigma^2}\right).$$

Proof. This also follows from Proposition 3.1. □

We note that in each model $\mathcal{M}_i$ the residual vector $e_{(i)}$ is a linear function of $e_r = M_1 y$, the residual vector in the restricted model. This is because $e_{(i)} = D_i y = D_i M_1 y$. All residuals (irrespective from which model) are therefore independent of $\hat{\beta}_{1r}$, the restricted estimator of β_1.

Finally, we present the covariances between the estimators from two competing models $\mathcal{M}_i$ and $\mathcal{M}_j$.

Proposition 3.4. *The estimators from models $\mathcal{M}_i$ and $\mathcal{M}_j$ are correlated with each other according to*

$$\begin{aligned}
cov(\hat{\beta}_{1(i)}, \hat{\beta}_{1(j)}) &= \sigma^2\left((X_1'X_1)^{-1} + QW_iW_jQ'\right),\\
cov(\hat{\beta}_{2(i)}, \hat{\beta}_{2(j)}) &= \sigma^2(X_2'M_1X_2)^{-1/2}W_iW_j(X_2'M_1X_2)^{-1/2},\\
cov(\hat{\beta}_{1(i)}, \hat{\beta}_{2(j)}) &= -\sigma^2QW_iW_j(X_2'M_1X_2)^{-1/2}.
\end{aligned}$$

Proof. Again, this follows from Proposition 3.1. □

As a consequence of Proposition 3.4 and assuming that $X_1'X_2 \neq 0$, we see that $\hat{\beta}_{2(j)}$ will be correlated with both $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}$, unless $W_iW_j = 0$.

4. WALS and the Equivalence Theorem

In the previous section, we derived the LS estimators for β_1 and β_2 in model $\mathcal{M}_i$, where β_2 is restricted by $S_i'\beta_2 = 0$. Since there are 2^{k_2} models, there are also 2^{k_2} different sets of estimators $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}$. The WALS estimators of β_1 and β_2 are, as the name suggests, weighted averages of the estimators of β_1 and β_2 over all models.

Definition 4.1. *The WALS estimators of β_1 and β_2 are*

$$b_1 = \sum_{i=1}^{2^{k_2}} \lambda_{(i)}\hat{\beta}_{1(i)}, \qquad b_2 = \sum_{i=1}^{2^{k_2}} \lambda_{(i)}\hat{\beta}_{2(i)},$$

where the sum is taken over all 2^{k_2} different models $\mathcal{M}_i$ obtained by setting a subset of the β_2's equal to zero.

The key question, of course, is how to define the model weights $\lambda_{(i)}$. (Here and in what follows we refer to the $\lambda_{(i)}$ as "model" weights to distinguish them from the "WALS" weights w_h to be defined later.) We shall impose the following restrictions.

Assumption 4.1. *The model weights $\lambda_{(i)}$ satisfy the following three regularity conditions:*

(R1) $0 \leq \lambda_{(i)} \leq 1$;
(R2) $\sum_i \lambda_{(i)} = 1$*; and*
(R3) $\lambda_{(i)} = \lambda_{(i)}(M_1 y)$.

The first two conditions simply state that the $\lambda_{(i)}$ are weights. Condition (R3), however, requires some justification. If σ^2 is known, then most or all pretest procedures will use statistics (such as t- and F-statistics) which depend on $\hat{\beta}_{2u}$ (that is, on $X_2' M_1 y$) only. If σ^2 is not known and is estimated by s_u^2 (the estimator of σ^2 in the unrestricted model), then all t- and F-statistics will depend on $(\hat{\beta}_{2u}, s_u^2)$. Now, since s_u^2 is a function of $M_1 y$, the pretest procedures will use statistics that depend on $M_1 y$ only. 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 $(\hat{\beta}_{2u}, s_u^2)$. However, they still depend only on $M_1 y$, because Propositions 3.1 and 3.3 imply that both $\hat{\beta}_{2(i)}$ and the residuals $e_{(i)}$ from model $\mathcal{M}_i$ are linear functions of $M_1 y$.

If we think of the model weights $\lambda_{(i)}$ as a measure of 'importance' of model $\mathcal{M}_i$ based on some diagnostic (say a t-statistic), then condition (R3) is satisfied. The regularity conditions on $\lambda_{(i)}$ thus appear to be reasonable and mild. They allow not only all standard pretest procedures, but also inequality-constrained least squares. Thus, Proposition 4.2 below explains the "surprising symmetry" found by Thomson and Schmidt (1982, p. 176).

Proposition 4.1. *Under regularity condition (R2), the WALS estimators take the form*

$$b_1 = \hat{\beta}_{1r} - Q W \hat{\theta}, \qquad b_2 = (X_2' M_1 X_2)^{-1/2} W \hat{\theta}, \tag{9}$$

where $W = \sum_i \lambda_{(i)} W_i$ is a symmetric random matrix (because the $\lambda_{(i)}$ are random) even though the W_i are nonrandom.

Proof. This follows by summing the expressions for $\hat{\beta}_{1(i)}$ and $\hat{\beta}_{2(i)}$ in (7) over all $i = 1, \ldots, 2^{k_2}$. Condition (R2) ensures that $\sum_i \lambda_{(i)} \hat{\beta}_{1r} = \hat{\beta}_{1r}$. □

We note that only condition (R2) is needed to obtain the WALS estimators. We also note that the dependence of the WALS estimators b_1 and b_2 on i is completely captured by the symmetric $k_2 \times k_2$ matrix W, which contains $k_2(k_2+1)/2$ essential elements, rather less than the 2^{k_2} different $\lambda_{(i)}$'s. As a consequence, in order to obtain the WALS estimates, we don't need to determine all the $\lambda_{(i)}$'s—it is sufficient to determine W. If W were diagonal, this would reduce our task even further, and this is precisely the purpose of the transformation introduced in Section 6.

Let us rewrite (9) as

$$\begin{aligned} b_1 - \beta_1 &= (X_1' X_1)^{-1}(X_1' \epsilon) - Q(W\hat{\theta} - \theta), \\ b_2 - \beta_2 &= (X_2' M_1 X_2)^{-1/2}(W\hat{\theta} - \theta). \end{aligned} \tag{10}$$

Suppose now that, in addition to (R2), regularity condition (R3) also holds, so that the $\lambda_{(i)}$ depend only on $M_1 y$. Then W will also depend only on $M_1 y$. Since $\hat{\theta}$ depends only on $M_1 y$ as well, condition (R3) guarantees that $W\hat{\theta}$ depends only on $M_1 y$ and is therefore independent of $X_1' \epsilon$. Based on these considerations, we prove the following key result.

Proposition 4.2. *(EQUIVALENCE THEOREM) If regularity conditions (R2) and (R3) on $\lambda_{(i)}$ are satisfied, then*

$$E(b_1) = \beta_1 - QE(W\hat{\theta} - \theta), \quad var(b_1) = \sigma^2(X_1'X_1)^{-1} + Qvar(W\hat{\theta})Q',$$

and hence

$$MSE(b_1) = \sigma^2(X_1'X_1)^{-1} + QMSE(W\hat{\theta})Q'.$$

Proof. This follows from the expressions in (10). Alternatively we have, since $\hat{\beta}_{1r}$ and $M_1 y$ are independent,

$$E(\hat{\beta}_{1r} \mid M_1 y) = E(\hat{\beta}_{1r}), \qquad var(\hat{\beta}_{1r} \mid M_1 y) = var(\hat{\beta}_{1r}),$$

so that

$$\begin{aligned} E(b_1 \mid M_1 y) &= E(\hat{\beta}_{1r} \mid M_1 y) - QE(W\hat{\theta} \mid M_1 y) \\ &= E(\hat{\beta}_{1r}) - QW\hat{\theta} = \beta_1 - Q(W\hat{\theta} - \theta) \end{aligned}$$

and

$$var(b_1 \mid M_1 y) = var(\hat{\beta}_{1r} \mid M_1 y) = var(\hat{\beta}_{1r}) = \sigma^2(X'X)^{-1}.$$

The unconditional mean and variance of b_1 and hence its mean squared error (MSE) follow. □

We shall assess the relative performance of estimators by comparing MSEs, that is, by assuming squared error loss. The properties of the complicated WALS estimator b_1 of β_1 thus depend critically on the properties of the less complicated estimator $W\hat{\theta}$ of θ. In particular, $MSE(b_1)$ is small whenever $MSE(W\hat{\theta})$ is small. 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 or, more accurately, on θ.

We can also interpret Proposition 4.2 in terms of β_1 *and* β_2. It follows from (10) that $MSE(b_1, b_2)$ can be written in the same form as $MSE(b_1)$, namely as $A + BMSE(W\hat{\theta})B'$ for some positive semidefinite matrix A and some matrix B, and hence the question how "good" b_1 and b_2 are as estimators of β_1 and β_2 depends completely on the question how 'good' $W\hat{\theta}$ is as an estimator of θ. This point is also emphasized by Clarke (2008).

Proposition 4.2 provides a nontrivial generalization of Theorem 2 in Magnus and Durbin (1999), using a simpler proof than in Magnus and Durbin (1999) and Danilov and Magnus (2004a). Extensions to cover, respectively, the cases of large-sample nonnormal errors and uncertainty about linear restrictions of the parameters β_1 and β_2 are provided in Zou *et al.* (2007) and Clarke (2008).

5. Interlude: Should Weights Lie between Zero and One?

Our results so far have not used condition (R1), requiring that the $\lambda_{(i)}$ lie between zero and one. Although we shall make this assumption later, this condition is not as obvious as it may appear at first. Following Magnus and Vasnev (2008), suppose we have two unbiased estimators of an unknown parameter μ:

$$\hat{\mu}_1 = \mu + \epsilon_1, \quad \hat{\mu}_2 = \mu + \epsilon_2,$$

where (ϵ_1, ϵ_2) follows a bivariate normal distribution with mean zero and known variance

$$var\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

If ϵ_1 and ϵ_2 are uncorrelated ($\sigma_{12} = 0$) and the two variances are equal, then we estimate μ by the unbiased minimum-variance estimator $\hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_2)/2$. If ϵ_1 and ϵ_2 are uncorrelated and the two variances are not equal, then the unbiased minimum-variance estimator of μ is

$$\hat{\mu} = \lambda\hat{\mu}_1 + (1-\lambda)\hat{\mu}_2, \quad \lambda = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

which is a weighted average of $\hat{\mu}_1$ and $\hat{\mu}_2$, where the weight λ lies between zero and one.

Next consider the case where ϵ_1 and ϵ_2 are correlated. Then we also obtain $\hat{\mu} = \lambda\hat{\mu}_1 + (1-\lambda)\hat{\mu}_2$, but now with

$$\lambda = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}.$$

We see that λ does *not* necessarily lie between zero and one. In particular, $\lambda < 0$ if and only if $\sigma_{12} > \sigma_2^2$, and $\lambda > 1$ if and only if $\sigma_{12} > \sigma_1^2$.

At first glance this may seem puzzling and unsatisfactory. At second glance, however, it becomes clear that this is the correct solution, and that we should not force our estimator to lie in-between the two underlying estimators.

To gain further insight let us consider two cases. First, the situation where the correlation is one ($\sigma_{12} = \sigma_1\sigma_2$) and $\sigma_1 \neq \sigma_2$. Then we have $\hat{\mu}_1 = \mu + \sigma_1\epsilon^*$ and $\hat{\mu}_2 = \mu + \sigma_2\epsilon^*$, where the common noise ϵ^* satisfies $\epsilon^* \sim \mathrm{N}(0, 1)$. In this case, $\hat{\mu}$ *must* lie outside the interval $(\hat{\mu}_1, \hat{\mu}_2)$. We simply solve the two equations in two unknowns (μ and ϵ^*) and find $\hat{\mu} = \lambda\hat{\mu}_1 + (1-\lambda)\hat{\mu}_2$ with

$$\lambda = \frac{\sigma_2}{\sigma_2 - \sigma_1},$$

The "weight" λ in this case is either larger than one (if $\sigma_1 < \sigma_2$) or smaller than zero (if $\sigma_1 > \sigma_2$).

Second, the situation where

$$\hat{\mu}_1 \sim \mathrm{N}(\mu, \sigma_1^2), \quad \hat{\mu}_2 = \hat{\mu}_1 + \epsilon_2$$

where ϵ_2 has mean zero and is distributed independently of $\hat{\mu}_1$. In this case $cov(\hat{\mu}_1, \hat{\mu}_2) = \sigma_1^2$ and $\hat{\mu} = \hat{\mu}_1$. The estimator $\hat{\mu}_1$ is a sufficient statistic for μ and the information contained in $\hat{\mu}_2$ is superfluous.

We conclude that – in the presence of correlation—"weights" may lie outside the (0, 1) interval. This result is not new, but it is little known because it is somewhat disturbing. Suppose the same king as in the Introduction now has two advisors, and that he consults both of them about next year's inflation. One predicts 2%, the other 4%. The first prediction has variance 1, the second has variance 4. If the two advisors were uncorrelated (they do not know each other and they base their forecasts on different data sets), then the king would weigh the two estimates, giving a higher weight to the first advisor because she is more accurate (has a lower variance). The answer then is 2.5%, which is in-between 2% and 4%, but closer to two than to four, as expected. But it is more likely that the advisors are correlated, that they do talk together, and that they use the same or similar data sets. If their correlation is 3/4 (which is not that high), then

$$\lambda = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}} = \frac{4 - 1.5}{1 + 4 - 2 \times 1.5} = 1.25,$$

and the king should therefore estimate next year's inflation by $1.25 \times 2\% - 0.25 \times 4\% = 1.5\%$, which lies outside the range indicated by the two advisors (2% and 4%). Now the king has a problem. He has a prediction which makes mathematical and statistical sense, but if he goes outside the range indicated by his advisors he will surely be heavily criticized. In practice, therefore, policy recommendations typically obey the boundaries specified by the advisors, and weights lie between zero and one. In accordance with all moving average techniques, we too submit to this practice.

6. A Semiorthogonal Transformation

The WALS procedure relies on a preliminary orthogonal transformation of the auxiliary regressors which greatly reduces the computational burden of the model-average estimator and has other advantages as well which will be discussed later.

We first scale the focus regressors X_1 by defining

$$Z_1 = X_1\Delta_1, \qquad \gamma_1 = \Delta_1^{-1}\beta_1, \tag{11}$$

where Δ_1 is a diagonal $k_1 \times k_1$ matrix with positive diagonal elements such that the diagonal elements of $Z_1'Z_1$ are all one. Notice that $Z_1\gamma_1 = X_1\beta_1$ and that

$$I_n - Z_1(Z_1'Z_1)^{-1}Z_1' = I_n - X_1\Delta_1(\Delta_1X_1'X_1\Delta_1)^{-1}\Delta_1X_1' = M_1.$$

Hence, scaling X_1 is completely harmless, and β_1 can always be recovered from γ_1 by $\beta_1 = \Delta_1\gamma_1$.

We next scale the auxiliary regressors X_2 by introducing a diagonal $k_2 \times k_2$ matrix Δ_2 with positive diagonal elements such that all diagonal elements of $\Delta_2X_2'M_1X_2\Delta_2$ are one. As a result, the original regressors X_1 and X_2 are scaled such that all diagonal elements of

$$(X_1\Delta_1)'(X_1\Delta_1) \quad \text{and} \quad (X_2\Delta_2)'M_1(X_2\Delta_2)$$

equal one. These scaling procedures were first proposed in De Luca and Magnus (2011) and stabilize both matrices so that inversion and eigenvalue routines become numerically more accurate.

For the auxiliary regressors we not only scale but also transform. Since the matrix $\Delta_2X_2'M_1X_2\Delta_2$ is positive definite, its eigenvalues are all positive and its eigenvectors are linearly independent. Defining the orthogonal $k_2 \times k_2$ matrix T (whose columns are the eigenvectors) and the diagonal $k_2 \times k_2$ matrix Ξ (containing the eigenvalues on the diagonal), we have

$$T'\Delta_2X_2'M_1X_2\Delta_2T = \Xi.$$

We then define

$$Z_2 = X_2\Delta_2T\Xi^{-1/2}, \qquad \gamma_2 = \Xi^{1/2}T'\Delta_2^{-1}\beta_2, \tag{12}$$

so that $Z_2\gamma_2 = X_2\beta_2$ and

$$Z_2'M_1Z_2 = \Xi^{-1/2}T'\Delta_2X_2'M_1X_2\Delta_2T\Xi^{-1/2} = I_{k_2}.$$

The effect of the scaling in X_1 is only for numerical stability; it has no effect on the WALS estimates. But the scaling in X_2 has two effects: numerical stability and scale-independence. This is because of the semiorthogonalization. Without preliminary scaling the WALS estimates would depend on the scaling of the auxiliary variables (unless $k_2 = 1$), because the orthogonal matrix T and the diagonal matrix Ξ depend on the scaling in a nontrivial (nonlinear) fashion. This dependence vanishes after preliminary scaling, which is obviously important in the interpretation of the WALS estimates. As with β_1, we can recover β_2 by $\beta_2 = \Delta_2T\Xi^{-1/2}\gamma_2$.

Since $X_1\beta_1 = Z_1\gamma_1$ and $X_2\beta_2 = Z_2\gamma_2$, the DGP given in (1) can also be written as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon. \tag{13}$$

The important difference between (1) and (13) is that $X_2'M_1X_2$ is positive definite but without any known structure, while $Z_2'M_1Z_2$ is constructed in such a way that it equals the identity matrix I_{k_2}.

7. Consequences of the Transformation

The fact that the matrix $M_1 Z_2$ is semiorthogonal so that $Z_2' M_1 Z_2 = I_{k_2}$ leads to important simplifications, which we list explicitly below. The matrices M_1 and Q become

$$M_1 = I_n - Z_1(Z_1'Z_1)^{-1}Z_1', \qquad Q = (Z_1'Z_1)^{-1}Z_1'Z_2,$$

and the LS estimators of γ_1 and γ_2 under the constraint $S_i'\gamma_2 = 0$ are now given by

$$\hat{\gamma}_{1(i)} = \hat{\gamma}_{1r} - Q W_i \hat{\gamma}_{2u}, \qquad \hat{\gamma}_{2(i)} = W_i \hat{\gamma}_{2u}, \tag{14}$$

where

$$\hat{\gamma}_{1r} = (Z_1'Z_1)^{-1}Z_1'y, \qquad \hat{\gamma}_{2u} = Z_2'M_1 y. \tag{15}$$

The idempotent matrix P_i reduces to

$$P_i = S_i(S_i'S_i)^{-1}S_i' = S_i S_i',$$

because S_i' is a selection matrix of the form $(I_{r_i} : 0)$ or a column-permutation thereof, so that $S_i'S_i = I_{r_i}$. Hence, P_i is a *diagonal* matrix with r_i ones and $k_2 - r_i$ zeros on the diagonal, and

$$W_i = I_{k_2} - S_i S_i'$$

is a *diagonal* matrix with $k_2 - r_i$ ones and r_i zeros on the diagonal.

The diagonality of W_i implies that the hth diagonal element of W_i is 0 if $\gamma_{2,h}$ (the hth component of γ_2) is constrained to be zero, and 1 otherwise. All models that include the hth column of Z_2 as a regressor will therefore have the *same* estimator of $\gamma_{2,h}$, irrespective which other γ_2's are estimated, namely the hth component of $\hat{\gamma}_{2u}$. Moreover,

$$\hat{\gamma}_{2u} \sim \mathrm{N}(\gamma_2, \sigma^2 I_{k_2}),$$

so that the k_2 components of $\hat{\gamma}_{2u}$ are independent.

The joint distribution of $\hat{\gamma}_{1(i)}$ and $\hat{\gamma}_{2(i)}$ is given by

$$\begin{pmatrix} \hat{\gamma}_{1(i)} \\ \hat{\gamma}_{2(i)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \gamma_1 + Q S_i S_i' \gamma_2 \\ W_i \gamma_2 \end{pmatrix}, \ \sigma^2 \begin{pmatrix} (Z_1'Z_1)^{-1} + Q W_i Q' & -Q W_i \\ -W_i Q' & W_i \end{pmatrix} \right),$$

and the residual vector is $e_{(i)} = y - Z_1\hat{\gamma}_{1(i)} - Z_2\hat{\gamma}_{2(i)} = D_i y$, where

$$D_i = M_1 - M_1 Z_2 W_i Z_2' M_1$$

is a symmetric idempotent matrix of rank $n - k + r_i$. The distribution of

$$s_{(i)}^2 = \frac{e_{(i)}'e_{(i)}}{n-k+r_i} = \frac{y' M_1 \left(I_n - Z_2 W_i Z_2'\right) M_1 y}{n-k+r_i} \tag{16}$$

is therefore given by

$$\frac{(n-k+r_i)\, s_{(i)}^2}{\sigma^2} \sim \chi^2\left(n-k+r_i, \ \frac{\gamma_2' S_i S_i' \gamma_2}{\sigma^2}\right),$$

The WALS estimators of γ_1 and γ_2 are

$$c_1 = \sum_{i=1}^{2^{k_2}} \lambda_{(i)} \hat{\gamma}_{1(i)} = \hat{\gamma}_{1r} - Q W \hat{\gamma}_{2u}, \qquad c_2 = \sum_{i=1}^{2^{k_2}} \lambda_{(i)} \hat{\gamma}_{2(i)} = W \hat{\gamma}_{2u}, \tag{17}$$

where

$$W = \sum_{i=1}^{2^{k_2}} \lambda_{(i)} W_i,$$

and the equivalence theorem implies that

$$MSE(c_1) = \sigma^2 (Z_1' Z_1)^{-1} + QMSE(W\hat{\gamma}_{2u})Q', \tag{18}$$

where W is a $k_2 \times k_2$ *diagonal* random matrix (random, because the $\lambda_{(i)}$ are random). Although the model space contains 2^{k_2} models, the computational burden of WALS is of the order k_2, because we need only consider the diagonal elements $w_1, \ldots, w_{k_2}$ of W, which are linear combinations of the model weights $\lambda_{(i)}$. We do not need to know these linear combinations explicitly, only the resulting WALS weights w_h ($h = 1, \ldots, k_2$).

It follows from the equivalence theorem (18) that the WALS estimator c_1 will be a "good" estimator of γ_1 (in the mean squared error sense) if and only if $W\hat{\gamma}_{2u}$ is a "good" estimator of γ_2. Now, W is diagonal and the elements of $\hat{\gamma}_{2u}$ are independent with $\hat{\gamma}_{2u,h} \sim \mathrm{N}(\gamma_{2,h}, \sigma^2)$. We want the diagonal elements w_h of W to lie between zero and one, so that the components of $W\hat{\gamma}_{2u}$ are shrinkage estimators. It is only at this stage that we require condition (R1) of Assumption 4.1, and in fact the condition is stronger than necessary. Some of the $\lambda_{(i)}$ may fall outside the (0, 1) interval as long as the k_2 required linear combinations (that is, the diagonal elements of $\sum_i \lambda_{(i)} W_i$) are all inside the (0, 1) interval. Thus, under semiorthogonalization and the full force of Assumption 4.1, it suffices to find the diagonal elements of W such that the shrinkage estimator $W\hat{\gamma}_{2u}$ is an "optimal" estimator of γ_2. Once we have found these elements, *the same* estimator will provide the optimal WALS estimator c_1 of γ_1 using (17).

Suppose that σ^2 is known (we discuss the unknown σ^2 case later). Then the relevant pretest procedures, and hence the $\lambda_{(i)}$, depend only on $Z_2' M_1 y$ as argued under Assumption 4.1. In that case we may strengthen condition (R3) that the $\lambda_{(i)}$ depend only on $M_1 y$ to the condition that they depend only on $Z_2' M_1 y$, that is, on $\hat{\gamma}_{2u}$. This is formalized in the following assumption.

Assumption 7.1. *The WALS weights w_h satisfy $w_h = w_h(\hat{\gamma}_{2u,h})$ for $h = 1, \ldots, k_2$.*

Not only does this seem reasonable, but is also has great practical advantages. In particular, since the $\{\hat{\gamma}_{2u,h}\}$ are independent, so are the $\{w_h \hat{\gamma}_{2u,h}\}$. Our k_2-dimensional problem thus reduces to k_2 (identical) one-dimensional problems: only using the information that $\hat{\gamma}_{2u,h} \sim \mathrm{N}(\gamma_{2,h}, \sigma^2)$ and assuming that σ^2 is known, find the best estimator of $\gamma_{2,h}$.

8. Estimating the Mean of a Univariate Normal Distribution from One Observation

Thus, motivated we address the seemingly trivial problem of estimating one parameter, say γ, given one observation, say x, generated by the normal $\mathrm{N}(\gamma, \sigma^2)$ distribution. Since we assume that the variance σ^2 is known, at least for the moment, there is no loss in generality by setting it equal to one.

We write our estimator as $m(x) = w(x)x$. The most obvious estimator of γ is $m(x) = x$ where $w(x) \equiv 1$. This estimator is unbiased, admissible, and minimax and its risk (which equals the mean squared error under squared error loss) is constant:

$$risk = E(x - \gamma)^2 = 1.$$

We call this the "usual" estimator. Another estimator is $m(x) = 0$ with $w(x) \equiv 0$. We call this the "silly" estimator. Its risk is

$$risk = E(0 - \gamma)^2 = \gamma^2.$$

If $|\gamma| < 1$ then the silly estimator has smaller risk, if $|\gamma| > 1$ then the usual estimator has smaller risk, and if $|\gamma| = 1$ then the two estimators have the same risk.

The equivalence theorem (Proposition 4.2) implies that associated with any estimator of γ in the above $N(\gamma, 1)$ problem there exists a unique estimator of β_1 in the regression problem defined in Section 2. For example, the unrestricted estimator $\hat{\beta}_{1u}$ corresponds to the usual estimator $m(x) = x$ of γ and the restricted estimator $\hat{\beta}_{1r}$ corresponds to the silly estimator $m(x) = 0$. Now, the usual estimator may make a lot of sense in the $N(\gamma, 1)$ context, but the unrestricted estimator $\hat{\beta}_{1u}$ makes less sense in the regression context, because it implies choosing $\hat{\beta}_{1u}$ whatever the values of the diagnostics associated with the auxiliary variables. The equivalence theorem thus shows that we have to reconsider the usefulness of the usual estimator also in the $N(\gamma, 1)$ context, and try and find an alternative to it.

There is no need for w to be constant. We can think of our estimator as a weighted average between the usual and the silly estimator, because

$$m(x) = w(x)x = w(x)x + (1 - w(x))0.$$

The larger is $|x|$ the larger should be w, so that more weight will be put on the usual estimator relative to the silly estimator. In fact, we impose the following regularity conditions on w.

Assumption 8.1. *The weight w is a real-valued function defined on $\mathbb{R}$ and satisfies:*

A. $0 \leq w(x) \leq 1$*;*
B. $w(-x) = w(x)$*;*
C. w *is nondecreasing on* $[0, \infty)$*;*
D. w *is continuous except possibly on a set of measure zero.*

Given these regularity conditions we obtain a lower bound for the risk.

Proposition 8.1. *If $x \sim N(\gamma, 1)$ and the weight w satisfies Assumption 8.1, then the risk $E(w(x)x - \gamma)^2$ has lower bound $\gamma^2/(1 + \gamma^2)$.*

Proof. This is proved in Magnus (2002, Theorem A.7). □

In search of a suitable w-function, we notice that any w satisfying Assumption 8.1 for which $w(0) = 0$ and $w(\infty) = 1$ can be viewed as a distribution function on $[0, \infty)$. Let us therefore consider a flexible three-parameter class of distribution functions, namely the reflected Burr class (Burr, 1942):

$$w(x) = 1 - (1 + (|x|/c)^{\alpha})^{-\delta} \qquad (c > 0,\ \alpha > 0,\ \delta > 0), \tag{19}$$

defined for $-\infty < x < \infty$. Given the reflected Burr class we minimize maximum regret, where regret is defined as

$$\begin{aligned} regret(\gamma; c, \alpha, \delta) &= risk(\gamma; c, \alpha, \delta) - \inf_{c,\delta,\alpha} risk(\gamma; c, \alpha, \delta) \\ &= risk(\gamma; c, \alpha, \delta) - \tfrac{\gamma^2}{1+\gamma^2}. \end{aligned}$$

Extensive optimization searches reveal that the minimax regret estimator is obtained along the path $\alpha\delta = 1$ when $\alpha \to \infty$. This gives

$$w(x) = \begin{cases} 0 & \text{if } |x| \leq c \\ 1 - c/|x| & \text{if } |x| > c, \end{cases}$$

and hence

$$m(x) = w(x)x = \begin{cases} x + c & \text{if } x < -c \\ 0 & \text{if } -c \leq x \leq c \\ x - c & \text{if } x > c. \end{cases} \tag{20}$$

This is the Burr estimator. The minimax regret solution is obtained for $c = 0.545$ with maximum regret equal to 0.3850. It seems therefore that we have solved the problem of finding the diagonal elements of W such that the shrinkage estimator $W\hat{\gamma}_{2u}$ is an "optimal" estimator of γ_2.

9. Enter Bayes: Neutrality and Robustness

We could stop here in our search for the "optimal" estimator of γ in the $N(\gamma, 1)$ problem. The Burr estimator, however, is not completely satisfactory. For example, it is "kinked," hence not differentiable and therefore inadmissible. Let us follow a different path, now along Bayesian lines. Our analysis so far has been strictly frequentist, but this section introduces a Bayesian element. The final product will thus contain both Bayesian and frequentist elements. The Bayesian solution will be close to the frequentist Burr solution, but it will be admissible and be based on a proper treatment of ignorance.

We start with the data,

$$x|\gamma \sim N(\gamma, 1),$$

which we combine with information from a prior $\pi(\gamma)$. This gives a posterior density $p(\gamma|x)$ of the form

$$p(\gamma|x) = \frac{\phi(x-\gamma)\pi(\gamma)}{\int_{-\infty}^{\infty} \phi(x-\gamma)\pi(\gamma)\,d\gamma},$$

where ϕ denotes the standard-normal density. The mean and variance of γ in the posterior distribution are denoted as $m(x)$ and $v(x)$, respectively.

If the prior is normal, then so is the posterior. In particular, if the prior distribution of γ is $N(0, \tau^2)$, then the mean and variance of γ in the posterior distribution are $m(x) = wx$ and $v(x) = w$, where $w = \tau^2/(\tau^2 + 1)$ is a constant. The normal prior, although convenient, is often considered inappropriate because the discrepancy between $m(x)$ and x does not vanish when x becomes large, but rather increases linearly without bound. In other words, the normal prior is not discounted when confronted with an observation with which it drastically disagrees. The normal prior is therefore not "robust" for the normal location problem.

We impose the following restrictions on the prior density.

Assumption 9.1. *The prior π is*

1. *symmetric around zero: $\pi(-\gamma) = \pi(\gamma)$ for all $\gamma > 0$;*
2. *positive and nonincreasing on $(0, \infty)$;*
3. *differentiable, except possibly at 0; and*
4. *$\omega(\gamma) = -\pi'(\gamma)/\pi(\gamma)$ has a limit (possibly ∞) as $\gamma \to \infty$.*

Assumptions (B1)–(B3) characterize the prior, allowing a nondifferentiable peak at zero. Assumption (B4) is a technical condition required in the proof of Proposition 9.1.

Robustness is formally defined as follows.

Definition 9.1. *A prior $\pi(\gamma)$ is said to be robust if the mean $m(x)$ in the posterior distribution based on this prior satisfies $x - m(x) \to 0$ as $x \to \infty$.*

To find out whether or not a prior is robust is not trivial, but the following proposition makes it easy.

Proposition 9.1. *Under Assumption 9.1, a prior π is robust if and only if $\omega(\gamma) \to 0$ as $\gamma \to \infty$.*

Proof. This is the main result of Kumar and Magnus (2013), and appears there as Theorem 1. □

For example, if the prior π follows a Student(k) distribution, then

$$\omega_k(\gamma) = \frac{-d\log\pi(\gamma)}{d\gamma} = \frac{(k+1)\gamma}{\gamma^2+k}.$$

We have $\lim_{\gamma\to\infty}\omega_k(\gamma) = 0$, but $\lim_{\gamma\to\infty}\lim_{k\to\infty}\omega_k(\gamma) = \infty$, because the sequence of functions $\{\omega_k\}$ ($k = 1, 2, \ldots$) is not uniformly convergent. This explains why the normal prior is not robust, while the Student prior is.

Let us consider a flexible three-parameter class of priors, namely the reflected generalized gamma distribution with density

$$\pi(\gamma) = \frac{qc^\delta}{2\Gamma(\delta)}|\gamma|^{-\alpha}e^{-c|\gamma|^q}, \tag{21}$$

where $-\infty < \gamma < \infty$, $c > 0$, $q > 0$, $0 \le \alpha < 1$ and $\delta = (1-\alpha)/q$. Special cases of (21) include the normal ($\alpha = 0$, $q = 2$ and $c = (2\sigma^2)^{-1}$), the Laplace ($\alpha = 0$ and $q = 1$), the reflected Weibull ($q = 1-\alpha$), and the Subbotin ($\alpha = 0$) distributions, among others. The ω-function takes the form

$$\omega(\gamma) = \frac{-d\log\pi(\gamma)}{d\gamma} = -\frac{\alpha}{\gamma} - cq\gamma^{q-1},$$

and hence robustness occurs if and only if $0 < q < 1$. We therefore restrict the parameter space to

$$c > 0, \quad 0 < q < 1, \quad 0 \le \alpha < 1. \tag{22}$$

In order to restrict the parameter space further we introduce the concept of neutrality.

Definition 9.2. *A prior $\pi(\gamma)$ is said to be neutral if the prior median of γ is zero and the prior median of $|\gamma|$ is one.*

The concept of neutrality attempts to capture the vague notion of ignorance in an explicit and applicable form. A prior is neutral when we are ignorant about the fact whether γ is positive or negative, and also about the fact whether $|\gamma|$ is larger or smaller than one. In other words, we require

$$\Pr(\gamma \le -1) = \Pr(-1 < \gamma \le 0) = \Pr(0 < \gamma \le 1) = \Pr(\gamma > 1) = 1/4.$$

The condition $\Pr(|\gamma| < 1) = \Pr(|\gamma| > 1) = 1/2$ corresponds precisely to ignorance about whether the usual or the silly estimator is better (has lower risk), and thus also ignorance about whether or not it is advisable to include the associated additional regressor.

For the reflected generalized gamma distribution defined in (21), neutrality occurs if and only if

$$\int_0^1 \gamma^{-\alpha}e^{-c\gamma^q}\,d\gamma = \frac{\Gamma(\delta)}{2qc^\delta}$$

If we define the (lower) incomplete gamma function as

$$\Gamma(s,z) = \frac{1}{\Gamma(s)}\int_0^z t^{s-1}e^{-t}dt$$

and notice that

$$\int_0^1 \gamma^{-\alpha}e^{-c\gamma^q}d\gamma = \frac{1}{qc^\delta}\int_0^c t^{\delta-1}e^{-t}dt = \frac{\Gamma(\delta,c)\Gamma(\delta)}{qc^\delta},$$

then an equivalent representation of the neutrality condition is

$$\Gamma(\delta, c) = 1/2. \tag{23}$$

Table 1. Minimax Regret Solutions for the Reflected Generalized Gamma, Reflected Weibull, Subbotin, and Laplace Priors

Prior	α	q	δ	c	γ	Regret
Gamma	0.2076	1	0.7924	0.4942	3.4410	0.4399
Weibull	0.1124	0.8876	1	0.6931	3.5620	0.4546
Subbotin	0	0.7995	1.2508	0.9377	3.6912	0.4697
Laplace	0	1	1	0.6931	4.9320	0.5127

In the case of the normal distribution we have $\delta = 1/2$ and $c = (2\sigma^2)^{-1}$, so that (23) holds for $\sigma^2 = 2.1981$. In the case of the Laplace and reflected Weibull distributions we have $\delta = 1$, and hence the neutrality condition becomes

$$\int_0^c e^{-t} dt = 1/2,$$

which leads to $c = \log 2$. More generally, the neutrality condition places a restriction on our class of priors by reducing the number of free parameters by one.

The minimax regret for our class of neutral and robust priors can then be written as

$$\inf_{c,q,\alpha} \sup_{\gamma} regret(\gamma; c, q, \alpha)$$
$$\text{subject to } \Gamma(\delta, c) = 1/2,$$

with the prior parameters satisfying the inequalities in (22).

Our minimax regret results are presented in Table 1. The Laplace prior is a special case of both the reflected Weibull and the Subbotin priors and is obtained by setting $q = 1$. It is neutral but not robust, and is included here as a benchmark. The reflected Weibull and Subbotin priors outperform the Laplace prior not only because they are robust while the Laplace prior is not but also because they have lower minimax regret. The reflected generalized gamma prior has a minimax solution on the boundary ($q = 1$) and this solution is not robust.

In Figure 1 we present the deviations $x - m(x)$ for the four minimax regret solutions. For the Laplace prior the deviation converges to $\log 2 = 0.6931$ and for the reflected generalized gamma it converges to $1/2$. Neither converges to zero in accordance to Proposition 9.1. For the reflected Weibull and Subbotin priors we see that $x - m(x)$ does converge to zero, confirming their robustness, also in accordance to Proposition 9.1.

The risk profiles of the reflected Weibull, Subbotin, and Burr estimators are presented in Figure 2 together with the minimum risk $\gamma^2/(1 + \gamma^2)$. The figure shows that the reflected Weibull and Subbotin estimators behave very similarly. Within the family of reflected generalized gamma distributions, as given in (21), the optimal neutral prior is obtained when $q = 1$, which is on the boundary and therefore not robust, because robustness requires that $0 < q < 1$. The optimal neutral *and* robust prior will therefore be obtained for q very close but not equal to one. Such a prior will then be robust but convergence of $x - m(x)$ to zero will be slow. This implies that we should choose q close but not too close to one.

Each of the three estimators in Figure 2 has its advantages and disadvantages. The Subbotin prior ($q = 0.80$) is "more robust" ($x - m(x)$ converges faster to zero) than the reflected Weibull ($q = 0.89$), but the reflected Weibull has lower minimax regret. The Burr estimator derived in the previous section has lower regret (0.3850) than any of the Bayesian estimators in the current section. However, it is

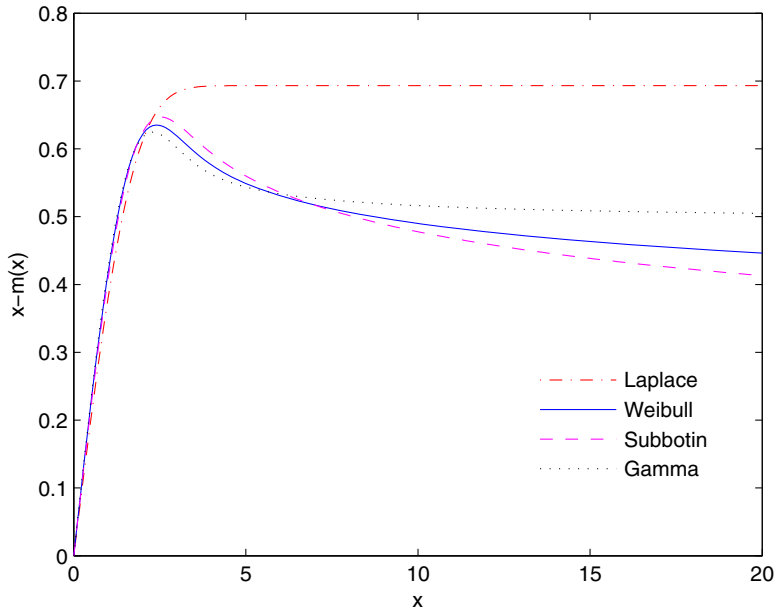

Figure 1. Deviations $x - m(x)$ for the Minimax Regret Solutions of the Reflected Generalized Gamma, Reflected Weibull, Subbotin, and Laplace Priors.

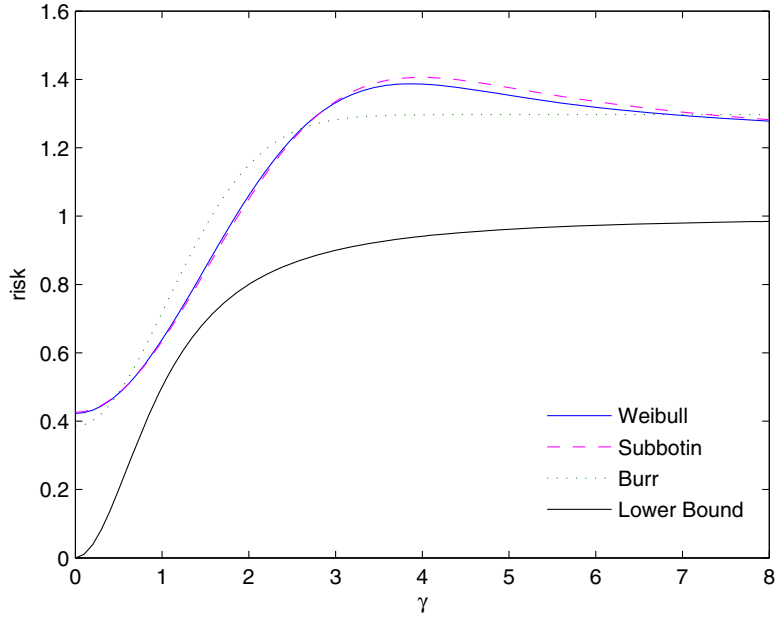

Figure 2. Risk of the Posterior Mean of the Reflected Weibull and Subbotin Priors, Compared to the Risk of the Burr Estimator.

inadmissible and does not have an interpretation in terms of ignorance. In our view the reflected Weibull prior,

$$\pi(\gamma) = \frac{qc}{2}|\gamma|^{-(1-q)}e^{-c|\gamma|^q} \qquad (c = \log 2,\ q = 0.8876), \tag{24}$$

offers the best compromise and we present it as our preferred prior. When we define

$$A_j(x) = \int_{-\infty}^{\infty}(x-\gamma)^j\phi(x-\gamma)\pi(\gamma)\,d\gamma \qquad (j = 0, 1, 2), \tag{25}$$

then the mean and variance in the posterior distribution are given by

$$m(x) = -\frac{A_1(x)}{A_0(x)} + x, \qquad v(x) = \frac{A_2(x)}{A_0(x)} - \left(\frac{A_1(x)}{A_0(x)}\right)^2. \tag{26}$$

10. The Effect of Estimating σ^2

So far we have assumed that the disturbance term in our linear model (1) is normally distributed with mean zero and variance $\sigma^2 I_n$, and that σ^2 is known. In fact, σ^2 is not known and we have to estimate it. The problem can be phrased in terms of the simple model of Section 8 and especially Section 9. There we considered the estimation of γ when we have one observation x from the univariate $N(\gamma, \sigma^2)$ distribution, assuming that σ^2 is known.

Suppose now that σ^2 is not known, but that we have an estimator s^2 of σ^2 which is distributed independently of x, such that $\nu s^2/\sigma^2$ follows a $\chi^2(\nu)$-distribution. This captures the essence of our problem. In Section 9 we obtained a weight function w such that $m(x) = w(x)x$ is an "optimal" estimator of γ, based on the reflected Weibull prior (24). This weight function was derived under the assumption that σ^2 is known. If σ^2 is not known we can ask two questions. First, how can we generalize the prior $\pi(\gamma)$ to a prior $\pi(\gamma, \sigma^2)$, thus obtaining a different posterior distribution than before, implying a different mean (and variance) in this distribution, and hence a different estimator of γ and a different risk profile? Second, if we use the same estimator as before (thus based on the known σ^2 situation), then what is the difference in risk when we replace σ^2 by s^2 and take the additional randomness of s^2 properly into account?

Both questions were analyzed in Danilov (2005), see also Yüksell *et al.* (2010), but in the context of the Laplace prior rather than the reflected Weibull. We conclude from Danilov's analysis that there is rapid convergence of the risk profiles to the known σ^2 case (where $\nu = \infty$) as we let ν approach ∞. For $\nu > 20$ the risk profiles practically coincide and even for small ν the difference is negligible.

We repeat one aspect of Danilov's analysis for the generalized Weibull. In Figure 3, we plot two risk functions. The first is the same risk function as the generalized Weibull in Figure 2, while the second uses the same formula but replaces σ^2 by s^2, thus allowing for the additional randomness caused by estimating σ^2. The risk now depends on the degrees of freedom ν and we choose a small value of ν ($\nu = 5$, as in Danilov (2005, figure 4)) in order to give randomness every chance to reveal itself. The figure shows that, even with such a small value of ν, the difference between the known and unknown σ^2 case is remarkably small, in fact negligible in practice. This result is not trivial or obvious. For example, when we draw the risk profiles of the pretest estimator, as in Danilov (2005, figure 3), we find large dependence on ν.

This concludes our discussion of the known versus unknown σ^2 case. The more general situation where the variance of the disturbances is not given by $\sigma^2 I_n$ but by $\sigma^2\Omega$, where Ω depends on some unknown parameters, is more complex and we shall discuss it when we discuss nonspherical disturbances in Section 12.1.

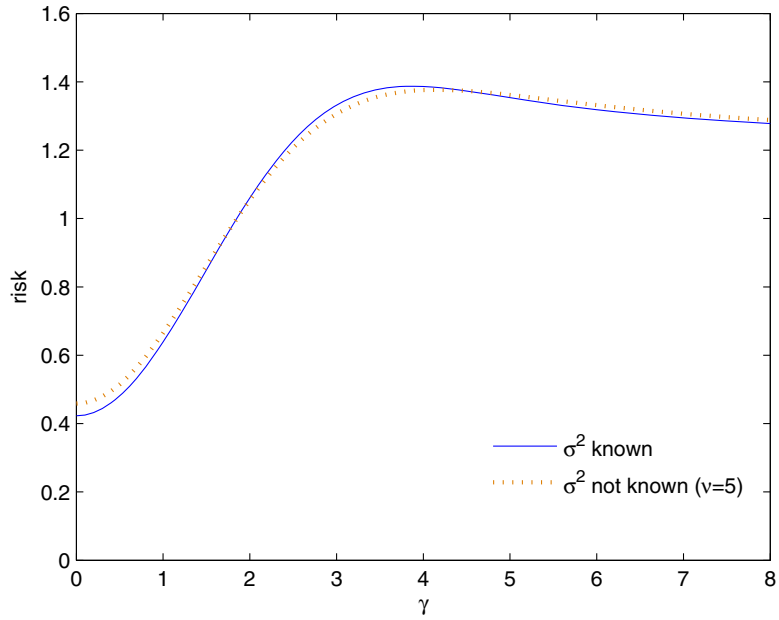

Figure 3. Risk of the Reflected Weibull Estimator: Known versus Unknown σ^2.

11. Putting It All Together

We summarize the WALS procedure in seven steps, as follows.

Step 1 (focus versus auxiliary). In the unrestricted model $y = X\beta + \epsilon$, determine which are the k_1 focus regressors (X_1) and which are the k_2 auxiliary regressors (X_2). This leads to (1):

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon.$$

Step 2 (scaling). Define the diagonal $k_1 \times k_1$ matrix Δ_1 whose jth diagonal element is given by

$$(\Delta_1)_{jj} = \frac{1}{\sqrt{(X_1'X_1)_{jj}}}.$$

Compute $M_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$ and define the diagonal $k_2 \times k_2$ matrix Δ_2 whose hth diagonal element is given by

$$(\Delta_2)_{hh} = \frac{1}{\sqrt{(X_2'M_1X_2)_{hh}}}.$$

As a result, all diagonal elements of

$$(X_1\Delta_1)'(X_1\Delta_1) \quad \text{and} \quad (X_2\Delta_2)'M_1(X_2\Delta_2)$$

are equal to one.

Step 3 (semiorthogonalization). Define

$$Z_1 = X_1\Delta_1, \qquad \gamma_1 = \Delta_1^{-1}\beta_1,$$

as in (11), so that $Z_1\gamma_1 = X_1\beta_1$. Next compute the orthogonal $k_2 \times k_2$ matrix T and the diagonal $k_2 \times k_2$ matrix Ξ such that

$$T'\Delta_2 X_2' M_1 X_2 \Delta_2 T = \Xi,$$

and define

$$Z_2 = X_2\Delta_2 T\,\Xi^{-1/2}, \qquad \gamma_2 = \Xi^{1/2}T'\Delta_2^{-1}\beta_2,$$

as in (12), so that $Z_2\gamma_2 = X_2\beta_2$ and $Z_2'M_1Z_2 = I_{k_2}$.

Step 4 (estimate γ_2 and σ^2 in the unrestricted model). Since $X_1\beta_1 = Z_1\gamma_1$ and $X_2\beta_2 = Z_2\gamma_2$, the DGP can also be written as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon,$$

as in (13). In this (unrestricted) model, compute the estimate of γ_2 as

$$\hat{\gamma}_{2u} = Z_2'M_1 y,$$

as in (15), and the estimate of σ^2 as

$$s_u^2 = \frac{y'M_1(I_n - Z_2Z_2')M_1 y}{n-k},$$

as in (16) with $r_i = 0$ and $W_i = I_{k_2}$. Set up the inputs for the Bayesian step by computing the vector of t-ratios

$$x = \frac{\hat{\gamma}_{2u}}{s_u} = \frac{\sqrt{n-k}\; Z_2'M_1 y}{\sqrt{y'M_1(I_{k_2} - Z_2Z_2')M_1 y}}.$$

Step 5 (compute the mean and variance in posterior distribution). Let ϕ denote the standard-normal density and define the reflected Weibull prior,

$$\pi(\gamma) = \frac{qc}{2}|\gamma|^{-(1-q)}e^{-c|\gamma|^q} \qquad (c = \log 2,\ q = 0.8876),$$

as in (24). For each of the k_2 components x_h of x compute

$$A_j(x_h) = \int_{-\infty}^{\infty}(x_h - \gamma)^j \phi(x_h - \gamma)\pi(\gamma)\,d\gamma \qquad (j = 0, 1, 2),$$

as in (25), and then the mean and variance in the posterior distribution:

$$m_h = m(x_h) = -\frac{A_1(x_h)}{A_0(x_h)} + x_h, \qquad v_h = v(x_h) = \frac{A_2(x_h)}{A_0(x_h)} - \left(\frac{A_1(x_h)}{A_0(x_h)}\right)^2,$$

as in (26). Compute $m = (m_1, \ldots, m_{k_2})'$ and $V = diag(v_1, \ldots, v_{k_2})$.

Step 6 (WALS estimates). Compute the WALS estimates of γ_1 and γ_2 as

$$c_1 = (Z_1'Z_1)^{-1}Z_1'(y - Z_2c_2), \qquad c_2 = s_u m.$$

The WALS estimates of the original parameters β_1 and β_2 are then given by

$$b_1 = \Delta_1 c_1, \qquad b_2 = \Delta_2 T\,\Xi^{-1/2}c_2.$$

Step 7 (WALS precisions). Letting $Q = (Z_1'Z_1)^{-1}Z_1'Z_2$, compute the variances of c_1 and c_2 as

$$var(c_1) = s_u^2(Z_1'Z_1)^{-1} + Q\,var(c_2)\,Q', \quad var(c_2) = s_u^2 V,$$

and the covariance as $cov(c_1, c_2) = -Qvar(c_2)$. The variances of b_1 and b_2 can then be computed as

$$var(b_1) = \Delta_1 var(c_1)\Delta_1, \quad var(b_2) = \Delta_2 T\,\Xi^{-1/2} var(c_2)\Xi^{-1/2}T'\Delta_2$$

and the covariance as $cov(b_1, b_2) = \Delta_1 cov(c_1, c_2)\Xi^{-1/2}T'\Delta_2$.

12. Extensions and Applications

The WALS procedure developed so far is designed for the estimation of linear regression models with i.i.d. disturbances, enabling the investigator to allow for both model uncertainty and estimation uncertainty in one joint procedure. Recent contributions have allowed extensions of this method, and in Section 12.1 we present the key ideas of four of these extensions. In Section 12.2, we discuss possible future extensions, and in Section 12.3 some applications. An important conclusion from the extensions is that the k_2-dimensional reduction, which occurs in the simple linear setting, continues to hold in most extensions by a method of transformation or linearization.

12.1 *Current Eextensions*

(a) *Nonspherical disturbances*. One constraint in the setup of WALS is that the disturbances are assumed to be independent and identically distributed with mean zero and variance σ^2. A more general setup would assume

$$\epsilon \sim \mathrm{N}(0, \sigma^2\Omega(\theta)),$$

where $\Omega(\theta)$ is a positive definite $n \times n$ matrix whose elements are functions of an m-dimensional unknown parameter vector $\theta = (\theta_1, \ldots, \theta_m)$ under the normalizing constraint $tr(\Omega(\theta)) = n$. This idea was developed in Magnus *et al.* (2011), mimicking the case where only σ^2 needs to be estimated.
If Ω were known we would transform model (1) to

$$\Omega^{-1/2}y = \Omega^{-1/2}X_1\beta_1 + \Omega^{-1/2}X_2\beta_2 + \Omega^{-1/2}\epsilon$$

and apply WALS to the transformed variables. Since Ω is not known, we estimate its parameters from the unrestricted model. This leads to the maximum likelihood (ML) estimator $\hat{\theta}$ of θ, through which we also obtain an estimator $\hat{\Omega} = \Omega(\hat{\theta})$. We then act as if $\hat{\Omega}$ is in fact the true rather than the estimated value of Ω, that is, we ignore the randomness in the estimation of the θ parameters.
The procedure can be justified by the analysis and results in Section 10, although there is one possibly important difference between estimating only σ^2 and estimating both σ^2 and θ. If $var(\epsilon) = \sigma^2 I_n$ then the estimators of β_1 and β_2 do not depend on σ^2, but if $var(\epsilon) = \sigma^2\Omega$ then the estimators of β_1 and β_2 do depend on Ω. The influence of the neglected randomness is therefore more complex, and so far it has not yet been explored.

(b) *Hierarchical WALS*. In practice the investigator has to decide not only which variables to include in the model, but also how to measure them. For example, one may wish to include "education" or "inflation" in the model, but then one has to decide as well how education and inflation are to be measured. This gives rise to two sources of model uncertainty: uncertainty about "groups" (such as eduction or inflation) and uncertainty about "variables" (different measurements of the group concept). This situation was recently analyzed by Magnus and Wang (2014) and is of interest, not only for its own sake, but also as an example where the number of regressors could exceed the number of observations, while estimation is still possible.

(c) *WALS estimation of generalized linear models*. A natural extension to the WALS methodology is to move away from the linearity assumption. Attempts to extend WALS to the wider class of generalized linear models (GLMs) were undertaken by Heumann and Grenke (2010) and De Luca *et al.* (2013). This class of models includes a variety of nonlinear models typically employed for discrete or categorical outcomes, such as logit, probit, and Poisson regression models.
In these models the dependence of the outcome on the regressors is modeled through a continuously differentiable and invertible function $h(\cdot)$, sometimes called the inverse link, such that $E(y|X) = h(X_1\beta_1 + X_2\beta_2)$. Because of model uncertainty about the k_2 columns of X_2, we consider 2^{k_2} possible models where the ith model $\mathcal{M}_i$ is again defined by a constraint of the form $S_i'\beta_2 = 0$. De Luca *et al.* (2013) show that, if we choose the unrestricted ML estimator $(\hat{\beta}_{1u}, \hat{\beta}_{2u})$ as the initial value, then a first-order approximation to the ML estimator of β_1 and β_2 under model $\mathcal{M}_i$ can be obtained. Moreover, after preliminary transformations of the outcome and the regressors, this (one-step) ML estimator closely resembles the constrained LS estimator of β_1 and β_2 discussed in Section 3.

(d) *WALS prediction*. Prediction model averaging with WALS is based on the linear DGP

$$\begin{pmatrix} y \\ y_f \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \\ X_{1f} & X_{2f} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon \\ \epsilon_f \end{pmatrix}, \tag{27}$$

where y and y_f are vectors of observations on the outcome, X_1 and X_{1f} are matrices of observations on the focus regressors, X_2 and X_{2f} are matrices of observations on the auxiliary regressors, β_1 and β_2 are vectors of focus and auxiliary parameters, and ϵ and ϵ_f are random vectors of unobservable disturbances. Observations are allowed to be correlated and we assume that

$$\begin{pmatrix} \epsilon \\ \epsilon_f \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & C_f' \\ C_f & \Omega_f \end{pmatrix} \right), \tag{28}$$

where the variance of (ϵ, ϵ_f) is a positive definite matrix whose component blocks Ω, C_f, and Ω_f are functions of a finite-dimensional unknown parameter vector θ.
After estimating β_1 and β_2 from the sample (y, X_1, X_2), we wish to predict the n_f (possibly future) values of y_f associated with the values of the regressors X_{1f} and X_{2f}. Because of model uncertainty, each of the k_2 columns of X_2 and X_{2f} can either be included or not included in the model and this gives rise to 2^{k_2} possible models. Unlike traditional prediction methods, model-average prediction procedures give predictors of y_f that take explicit account of both model and error uncertainty. The additional theory required for WALS prediction was recently developed in Magnus *et al.* (2014).

12.2 *Possible Future Extensions*

In addition to the four extensions discussed in the previous subsection, there are many other possible extensions of WALS, and much work is still required to fill these gaps. Below we list five extensions that seem most important to us, but it is easy to think of more extensions, for example panel data, big data ($k \gg n$), or semi- or nonparametric settings.

(a) *Endogeneity*. The standard setup requires that the regressors are (weakly) exogenous. If we allow endogenous regressors then we need a WALS version of instrumental variables or two-stage least squares.

(b) *Role of normality*. The normality assumption plays an important role in the development of WALS, which goes beyond the specification of the first two moments. At several points in the development we need the fact that if two random variables are uncorrelated, then functions of these random variables are also uncorrelated. This is generally not true, but under normality it is. An extension

to nonnormal errors is therefore far from trivial. Zou *et al.* (2007) provide an extension of the equivalence theorem for large-sample nonnormal errors and De Luca *et al.* (2013) generalize WALS to GLMs where the density of y is assumed to belong to the linear exponential family of density functions. While these extensions are useful, they do not yet provide a full generalization to nonnormality.

(c) *Multivariate models*. Extensions from univariate to multivariate models should be feasible within the WALS framework, but so far this has not been investigated. These extensions would open the way to a larger variety of models, such as seemingly unrelated regression equations (SURE), and ordered, multinomial, and conditional logit and probit models.

(d) *Excluding subsets of parameters*. In the standard WALS setup we allow all 2^{k_2} models to play a part, where the ith model is identified by a linear restriction of the form $S_i'\beta_2 = 0$. It can happen that if one model is excluded then another model is automatically excluded as well or if one auxiliary regressor appears then this excludes a subset of other auxiliary regressors. The hierarchical procedure in Section 12.1 is an example of the latter situation. A general theory of subset selection allowing for linkage between restrictions would be of great practical importance.

(e) *Nonnested models*. Our models are all nested within one (the largest) model, which is also the DGP. A general theory of model selection should allow for the fact that models may not be nested.

12.3 *Applications*

Model averaging techniques are typically applied to growth empirics, where the large number of growth determinants and the limited number of observations available at the national level expose the regressions to a high degree of model uncertainty. The standard WALS procedure, introduced by Magnus *et al.* (2010), also contains an application to growth empirics. The WALS estimates obtained in Magnus *et al.* (2010) are compared to traditional pretest estimates, Bayesian model-average (BMA) estimates, and Bayesian averaging of classical estimates, using the data set analyzed by Sala-i-Martin *et al.* (2004).

The hierarchical WALS estimator discussed in Section 12.1 was applied to growth empirics by Magnus and Wang (2014) in order to take explicit account of the uncertainty due to the choice of the growth determinants (that is, concepts or groups of variables such as education) and the choice of the explanatory variables that can be employed as alternative proxy measures of the same growth determinant (that is, precisely defined variables such as enrollment rate, school years, and share of public education spending). Poghosyan and Magnus (2012) applied WALS to estimate and forecast factor-based dynamic models of real GDP growth and inflation in the Armenian economy. Additional comparisons between WALS, standard BMA, and a modified version of the Mallows model-average (MMA) estimator of Hansen (2007) are provided by Amini and Parmeter (2012) using data from three earlier studies on growth empirics. Similar comparisons between WALS and other frequentist-based model-average estimators can be found in Liu (2014).

The WALS procedure has also been successfully applied in other fields, outside growth empirics. Among these, Wan and Zhang (2009) applied WALS to study the effect of recreation and tourism development on a range of socioeconomic indicators in rural U.S. counties. Liski *et al.* (2010) used WALS to investigate medical care costs of hip fracture treatments in hospital districts of Finland. Magnus *et al.* (2011) provided an application of WALS in the context of a hedonic housing price model with heteroskedastic disturbances using data from the Hong Kong real estate market; Seya and Tsutsumi (2012) extended this analysis to WALS estimation of a hedonic land price model with spatially dependent data from the Tokyo metropolitan area; and Seya *et al.* (2014) provided Monte Carlo simulations on the combined use of WALS and principal components to jointly address problems of model uncertainty and multicollinearity in spatial lag-error models.

Dardanoni *et al.* (2011, 2012) exploited WALS for handling the bias-precision trade-off that arises in the estimation of linear regression models with missing covariate values replaced by imputations. Osterloh (2012) applied WALS to assess the robustness of political environment as determinant of the economic performance in OECD countries. Finally, Cook *et al.* (2013) used WALS to study the determinants of various capital structure measures in US corporations between 1980 and 2007. In contrast to the growth empirics literature, where the number of regressors is large, some of the above studies show that WALS can be used as a general method of estimation, also when the number of regressors is small.

13. WALS Compared to other Model-Average Estimators

The method of WALS surveyed in this paper is one of several methods of model averaging (more accurately, estimator averaging). In this section we provide a brief overview of this rapidly expanding literature, emphasizing the advantages and weaknesses of WALS relative to these alternative model-average procedures.

13.1 *Bayesian Model Averaging*

BMA estimators compute weighted averages of the conditional estimates over all possible models using posterior model probabilities as weights in order to reflect the confidence in each model based on prior beliefs and the observed data. By Bayes theorem, the posterior probability of model $\mathcal{M}_i$ is obtained as

$$\lambda_{(i)} = p(\mathcal{M}_i \mid y) = \frac{p(\mathcal{M}_i)\, p(y \mid \mathcal{M}_i)}{\sum_i p(\mathcal{M}_i)\, p(y \mid \mathcal{M}_i)} \qquad (i = 1, \ldots, 2^{k_2}), \tag{29}$$

where $p(\mathcal{M}_i)$ is the prior probability that $\mathcal{M}_i$ is the true model,

$$p(y \mid \mathcal{M}_i) = \int p(y \mid \theta_i, \mathcal{M}_i)\, p(\theta_i \mid \mathcal{M}_i)\, d\theta_i \tag{30}$$

is the marginal likelihood of model $\mathcal{M}_i$, θ_i the vector of its parameters, $p(y \mid \theta_i, \mathcal{M}_i)$ its likelihood, and $p(\theta_i \mid \mathcal{M}_i)$ the prior density of θ_i under model $\mathcal{M}_i$. Thus, contrary to WALS which uses priors only on the vector of t-ratios $x = \hat{\gamma}_{2u}/s_u^2$, BMA estimators require two types of priors: on the model space and on the parameters of each model.

As discussed in the reviews by Hoeting *et al.* (1999), Clyde and George (2004), and Moral-Benito (2013), the choice of uninformative priors is one of the most challenging aspects of BMA. For the prior on the model space, the bulk of the BMA literature uses a uniform prior which assigns equal probability to each model. Although this is a reasonable choice when there is little prior information about the relative plausibility of the models considered, the uniform prior has been criticized 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; see, for example, Durlauf *et al.* (2008). For the prior on the parameter space, the major problem is that improper priors on all parameters would result in ill-defined Bayes factors. A widely used strategy to deal with this issue consists of using a hierarchical prior structure that involves improper priors on the parameters that are common to all models and proper priors on the remaining parameters. For the latter, most BMA estimators rely on Zellner's (1986) g-prior, which is a normal prior with mean zero and variance

$$var(\beta_2 \mid \mathcal{M}_i) = \frac{\sigma^2}{g} S_i (S_i' X_2' M_1 X_2 S_i)^{-1} S_i', \tag{31}$$

where $g > 0$ is a scalar hyperparameter that reflects how much importance is given to prior beliefs. This prior structure is attractive since it only requires the elicitation of the scalar parameter g and, in linear

regression setups, leads to closed-form expressions of the posterior model probabilities. For the choice of the hyperparameter g, various options are available including the unit information prior $g = 1/n$ (Kass and Wasserman, 1995), the risk inflation criterion $g = 1/k_2^2$ (Foster and George, 1994), and the benchmark prior $g = 1/\max(n, k_2^2)$ (Fernández *et al.*, 2001). For additional references and recent developments on the choice of priors in BMA, the reader is referred to Eicher *et al.* (2011) and Ley and Steel (2009, 2012).

Another well-known practical issue that plagues standard BMA procedures is how to handle model spaces of large dimensions. A widely used empirical strategy consists of using approximate estimates that consider only a suitable subset of models supported by the data. The subset of models to be investigated can be identified either by deterministic search methods such as Occam's window (Madigan and Raftery, 1994) and the leaps and bounds algorithm (Furnival and Wilson, 1974), or by stochastic search methods based on Markov chain Monte Carlo (MCMC) techniques; see Garcia-Donato and Martinez-Beneito (2013) for a recent review.

Several BMA estimators have been developed in the context of the linear regression model, but extensions to more general regression setups are also available. For example, BMA estimation of GLMs has been considered in Raftery (1996), Clyde (2000), and Czado and Raftery (2006). More recently, Jordan and Lenkoski (2012) and Eicher *et al.* (2012) focused on BMA estimation of models for truncated and censored data; Crespo Cuaresma and Feldkircher (2012) applied BMA in the presence of spatial autocorrelation using spatial filtering; Koop *et al.* (2012), Karl and Lenkoski (2012), and Lenkoski *et al.* (2013) developed BMA methodologies to jointly address model uncertainty and endogeneity; and Chen *et al.* (2009), Moral-Benito (2012), and McCormick *et al.* (2012) proposed BMA estimators for panel data models. Notice that, outside the classical linear regression model with conjugate priors, the marginal likelihoods in (30) do not usually have analytic closed-form expressions. Approximations to either the marginal likelihoods or the posterior model probabilities are therefore needed. Such approximations can be obtained through the Laplace method for integrals (Tierney and Kadane, 1986), the output of some MCMC method (Han and Carlin, 2001; Ghosh and Clyde, 2011), or a suitable combination of these two methods (DiCiccio *et al.*, 1997; Lewis and Raftery, 1997).

13.2 *Frequentist Model Averaging*

Frequentist model averaging (FMA) differs from BMA and WALS in that it avoids the need of priors elicitation, because FMA uses model weights that are totally determined by the data on the basis of some diagnostic criterion. Useful overviews of this approach can be found in Claeskens and Hjort (2008) and Wang *et al.* (2009). The frequentist perspective to model averaging is more recent than the Bayesian perspective, and significant progress has recently been made in developing optimal data-driven weighting schemes and investigating the properties of the resulting estimators. For example, Buckland *et al.* (1997) suggested mixing models with weights based on smoothed Akaike information criterion or Bayesian information criterion scores; Yang (2001, 2003) proposed a frequentist-based adaptive regression mixing method that allows the combination of estimators from different estimation procedures; Hjort and Claeskens (2003) introduced a likelihood-based local misspecification framework to analyze the asymptotic distribution of model-average estimators; Hansen (2007) and Wan *et al.* (2010) developed least-squares model-average estimators with weights selected by minimizing the Mallows criterion; and Liang *et al.* (2011) proposed a weighting procedure which minimizes an unbiased estimator of the mean squared error of the FMA estimator in finite samples.

Most of the available distributional results of FMA estimators are based on large-sample approximations and they are typically established under the assumptions of the local misspecification framework (Hjort and Claeskens, 2003). The work of Pötscher (2006), who investigated the finite-sample and asymptotic distributions of a special case of the FMA estimator discussed in Leung and Barron (2006), provides a helpful guide for further research on these topics.

As for BMA, exact FMA estimation can be computationally very demanding when the model space is large. This issue is typically addressed through a preliminary model-screening step that removes the poorest-performing models before combining the conditional estimates into an unconditional FMA estimate. Applications of this preliminary model screening step can be found in Yuan and Yang (2005), Claeskens *et al.* (2006), Wan *et al.* (2014), and Zhang *et al.* (2013a), among others. The computation of exact FMA estimates in growth applications with large model spaces was first attempted in Amini and Parmeter (2012) by introducing an operational version of the model-average estimator of Hansen (2007), using the same semiorthogonal transformations as adopted in WALS.

Extensions of FMA estimators to more general regression setups cover a variety of models such as logistic regression models (Claeskens *et al.*, 2006); models for survival analysis (Hjort and Claeskens, 2006); generalized additive partial linear models (Zhang and Liang, 2011); Tobit models (Zhang *et al.*, 2012); multinomial and ordered logit models (Wan *et al.*, 2014); and linear mixed-effects models (Zhang *et al.*, 2014). For linear models with nonsperical disturbances, Hansen and Racine (2012) and Zhang *et al.* (2013b) developed jackknife model-average estimators which are asymptotically optimal under heteroskedastic and serially correlated errors, while Liu and Okui (2013) and Liu *et al.* (2013) extended the MMA estimator of Hansen (2007) to models with heteroskedastic errors. In the context of prediction, Hansen (2008) extended the idea of Mallows model averaging to forecast combinations, while Yang (2004), Zou and Yang (2004), and Zhang *et al.* (2013a) developed combining forecasting procedures for time-series models. Finally, Schomaker *et al.* (2010) examined the properties of various FMA estimators in the presence of missing data and Kuersteiner and Okui (2010) used the MMA estimator of Hansen (2007) to construct optimal instruments in the context of linear models with endogenous regressors.

13.3 *Intermediate Position of WALS*

The WALS procedure surveyed in this paper is a Bayesian combination of frequentist estimators. The parameters of each model are estimated by constrained least squares, hence frequentist. However, after implementing a semiorthogonal transformation to the auxiliary regressors, the weighting scheme is developed on the basis of a Bayesian approach in order to obtain desirable theoretical properties such as admissibility and a proper treatment of ignorance. The final result is a model-average estimator that assumes an intermediate position between strict BMA and strict FMA estimators.

WALS is closely related to the FMA estimator proposed by Liang *et al.* (2011). The model setup, the assumptions, and the estimator are the same, but Liang *et al.* choose the weights to minimize an unbiased estimator of the mean squared error of the FMA (or WALS) estimator rather than of the mean squared error itself.

WALS is conceptually also close to BMA. The assumption that the data are normally distributed is the same, and the treatment of noninformative priors on the model space, the focus parameters β_1 and the error variance σ^2 is essentially the same. The main difference between the two model-average procedures lies in the prior treatment of the auxiliary parameters β_2. In WALS we write β_2 as $\beta_2 = \sigma \Delta_2 T \Xi^{-1/2} \gamma$, where $\gamma = \gamma_2/\sigma$, and assume that the k_2 components of γ are i.i.d. according to a reflected Weibull distribution

$$\pi(\gamma_h) = \frac{qc}{2} |\gamma_h|^{-(1-q)} \exp(-c|\gamma_h|^q)$$

with $c = \log 2$ and $q = 0.8876$. This implies that each γ_h is symmetrically distributed around zero, that the median of γ_h^2 is one, and that the variance of γ_h is $\sigma_\gamma^2 = \Gamma((q+2)/q)/c^2$. As argued in Section 9, this choice of prior moments is based on the concept of neutrality which attempts to formalize the vague notion of ignorance in an explicit and applicable form, and on other theoretical considerations related to robustness and optimality in the minimax regret sense.

The resulting prior mean of β_2 is zero and its prior variance is given by

$$var(\beta_2) = \sigma^2 \sigma_\gamma^2 \Delta_2 T \Xi^{-1} T' \Delta_2 = \frac{\sigma^2}{c^2/\Gamma((q+2)/q)} (X_2' M_1 X_2)^{-1}. \quad (32)$$

A comparison of (31) and (32) shows that the prior moments adopted in BMA and WALS are in fact closely related, and suggests in addition a new value for g in BMA applications, namely $g = c^2/\Gamma((q+2)/q) = 0.1878$.

A key feature of WALS is that the choice of prior is not *ad hoc* (like in BMA) but theoretically based. A second key feature of WALS provides a practical rather than a theoretical advantage over both BMA and FMA, namely that, even though there are 2^{k_2} possible models to consider, the computational burden of this model-average estimator is reduced to the order k_2 by the semiorthogonal transformation of the auxiliary regressors discussed in Section 6. Thus, while standard BMA and FMA estimators require sophisticated approximation algorithms to explore small subsets of the model space, WALS provides exact model-average estimates of the parameters of interest in negligible computing time. This computational advantage is likely to play an important role in a variety of empirical applications where the estimation of all models is impossible.

Both from a theoretical and a practical viewpoint WALS has many attractive features. Of course, the semiorthogonal transformation—which leads to the simplifications which WALS requires—also has a cost, both theoretically and practically. From a theoretical viewpoint, some of our assumptions are not made directly on the auxiliary parameters β_2 but rather on the transformed parameters γ_2, whose components are linear combinations of the components of β_2. For example, the notion of ignorance is defined in terms of γ_2, not in terms of β_2. This can be defended because both sets are auxiliary parameters, and if we are ignorant about one set we are equally ignorant about the other. From a more practical viewpoint, the flexibility of the WALS procedure is bounded, although we do not know yet precisely what the bounds are. Consider, for example, certain forms of model uncertainty that require mixing nonnested models such as the choice of the link function in a GLM setup. While these forms of model uncertainty can be handled by standard model-average procedures (Czado and Raftery, 2006), the generalization of WALS along these directions appears to be difficult.

Finally we emphasize (again) that WALS is a model-average procedure, not a model-selection procedure. At the end we cannot and do not want to answer the question: which model is best? This brings with it certain restrictions. For example, WALS cannot handle jointness (Ley and Steel, 2007; Doppelhofer and Weeks, 2009). The concept of jointness refers to the dependence between explanatory variables in the posterior distribution, and available measures of jointness depend on posterior inclusion probabilities of the explanatory variables, which WALS does not provide.

14. Software for WALS

User-friendly packages for WALS estimation of linear regression models with i.i.d. errors are available both in `MATLAB` and `Stata`. The latest 2.0 version of both packages—together with documentation, examples, and supplementary material—can be downloaded free of charge from the website http://www.janmagnus.nl/items/WALS.pdf.

Version 2.0 differs from earlier versions of the WALS packages in two important respects. First, we have enlarged the set of priors available for WALS estimation to three possible distributions: Weibull (the default), Subbotin, and Laplace. Parameters of all these prior distributions are always fixed to their minimax regret solutions under the neutrality condition.

Second, because of differences in the integration routines available in `MATLAB` and `Stata`, we have implemented two alternative strategies for computing the mean and variance in the posterior distribution. By default, the integrals involved in the computation of these posterior moments are solved numerically

by high-order global adaptive quadrature in the `MATLAB` package and by Gauss-Laguerre quadrature with 100 data points in the `Stata` package (Cheney and Kincaid, 2008). Even though these integration routines seem to produce very similar results in our examples, we cannot guarantee that this is true in general. For this reason, we introduced the `postmoments` option which interpolates $m(x)$ and $v(x)$ between the moments of the nearest x-values contained in precompiled tables of posterior means and variances under the Weibull and Subbotin priors at given x-values in the interval $[0, 100]$ with step 0.01. These tables have been computed in `MATLAB` using a high degree of accuracy. For $x > 100$, posterior moments are approximated by $m(100)$ and $v(100)$. For the Laplace prior, the `postmoments` option is not active since moments of the resulting posterior distribution can be computed accurately in any statistical software.

No user-friendly packages are currently available for generalizations of the WALS procedure to more general regression models. However, as discussed in Section 12, most of these generalizations can be easily implemented either by applying standard WALS estimates to suitable transformations of the data, or by computing simple weighted averages of standard WALS estimates.

Acknowledgments

This survey is, in essence, the content of a series of lectures given at the Einaudi Institute for Economics and Finance (EIEF) in Rome and at the Dipartimento di Scienze Economiche, Aziendali e Statistiche (SEAS) in Palermo on the subject "pretesting and model averaging" in April and November 2013. We are grateful to the participants for their careful reading and constructive comments. The editor and three referees provided embarrassingly positive and insightful suggestions, which helped to sharpen the message of the paper.

References

Amini, S.M. and Parmeter, C.F. (2012) Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics* 27: 870–876.

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997) Model selection: An integral part of inference. *Biometrics* 53: 603–618.

Burr, I.W. (1942) Cumulative frequency functions. *Annals of Mathematical Statistics* 13: 215–232.

Chen, H., Mirestean, A. and Tsangarides, C. (2009) Limited information Bayesian model averaging for dynamic panels with short time periods. Working Paper 2009-74, IMF.

Cheney, W. and Kincaid, D. (2008) *Numerical Mathematics and Computing* (6th ed). Belmont, CA: Thomson Brooks/Cole.

Claeskens, G., Croux, C. and van Kerckhoven, J. (2006) Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* 62: 972–979.

Claeskens, G. and Hjort, N.L. (2008) *Model Selection and Model Averaging*. New York: Cambridge University Press.

Clarke, J.A. (2008) On weighted estimation in linear regression in the presence of parameter uncertainty. *Economics Letters* 100: 1–3.

Clyde, M.A. (2000) Model uncertainty and health effect studies for particulate matter. *Environmetrics* 11: 745–763.

Clyde, M.A. and George, E.I. (2004) Model uncertainty. *Statistical Science* 19: 81–94.

Cook, D.O., Keefe, M.O.C. and Kieschnick, R. (2013) The implications of capital structure measurement for capital structure research. Mimeo.

Crespo Cuaresma, J. and Feldkircher, M. (2012) Spatial filtering, model averaging and the speed of income convergence in Europe. *Journal of Applied Econometrics* 28: 720–741.

Czado, C. and Raftery, A.E. (2006) Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors. *Statistical Papers* 47: 419–442.

Danilov, D. (2005) Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal* 8: 277–291.

Danilov, D. and Magnus, J.R. (2004a) On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122: 27–46.
Danilov, D. and Magnus, J.R. (2004b) Forecast accuracy after pretesting with an application to the stock market. *Journal of Forecasting* 23: 251–274.
Dardanoni, V., Modica, S. and Peracchi, F. (2011) Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics* 162: 362–368.
Dardanoni, V., De Luca, G., Modica, S. and Peracchi, F. (2012) A generalized missing-indicator approach to regression with imputed covariates. *Stata Journal* 12: 575–604.
De Luca, G. and Magnus, J.R. (2011) Bayesian model averaging and weighted average least squares: Equivariance, stability and numerical issues. *Stata Journal* 11: 518–544.
De Luca, G., Magnus, J.R. and Peracchi, F. (2013) Nonlinear weighted average least squares. Mimeo.
DiCiccio, T.J., Kass, R.E., Raftery, A.E. and Wasserman, L. (1997) Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* 92: 903–915.
Doppelhofer, G. and Weeks, M. (2009) Jointness of growth determinants. *Journal of Applied Econometrics* 24: 209–244.
Durlauf, S.N., Kourtellos, A. and Tan, C.M. (2008) Are any growth theories robust? *Economic Journal* 118: 329–346.
Eicher, T.S., Papageorgiou, C. and Raftery, A.E. (2011) Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics* 26: 30–55.
Eicher, T.S., Helfman, L. and Lenkoski, A. (2012) Robust FDI determinants: Bayesian model averaging in the presence of selection bias. *Journal of Macroeconomics* 34: 637–651.
Fernández, C., Ley, E. and Steel, M.F.J. (2001) Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100: 381–427.
Foster, D.P. and George, E.I. (1994) The risk inflation criterion for multiple regression. *Annals of Statistics* 22: 1947–1975.
Furnival, G.M. and Wilson, R.W. (1974) Regression by leaps and bounds. *Technometrics* 16: 499–511.
Garcia-Donato, G. and Martinez-Beneito, M.A. (2013) On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association* 108: 340–352.
Ghosh, J. and Clyde, M.A. (2011) Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association* 106: 1041–1052.
Giles, J.A. and Giles, D.E.A. (1993) Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys* 7: 145–197.
Han, C. and Carlin, B.P. (2001) Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association* 96: 1122–1132.
Hansen, B.E. (2007) Least squares model averaging. *Econometrica* 75: 1175–1189.
Hansen, B.E. (2008) Least-squares forecast averaging. *Journal of Econometrics* 146: 342–350.
Hansen, B.E. and Racine, J.S. (2012) Jackknife model averaging. *Journal of Econometrics* 167: 38–46.
Heumann, C. and Grenke, M. (2010) An efficient model averaging procedure for logistic regression models using a Bayesian estimator with Laplace prior. In T. Kneib and G. Tutz (eds), *Statistical Modelling and Regression Structures* (pp. 79–90). Heidelberg: Physica-Verlag.
Hjort, N.L. and Claeskens, G. (2003) Frequentist model averaging estimators. *Journal of the American Statistical Association* 98: 879–899.
Hjort, N.L. and Claeskens, G. (2006) Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association* 101: 1449–1464.
Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: A tutorial. *Statistical Science* 14: 382–417.
Jordan, A. and Lenkoski, A. (2012) Tobit Bayesian model averaging and the determinants of foreign direct investment. Mimeo.
Judge, G.G. and Bock, M.E. (1978) *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.
Karl, A. and Lenkoski, A. (2012) Instrumental variable Bayesian model averaging via conditional Bayes factors. Mimeo.

Kass, R.E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90: 928–934.
Koop, G., Leon-Gonzalez, R. and Strachan, R. (2012) Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics* 171: 237–250.
Kuersteiner, G. and Okui, R. (2010) Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78: 697–718.
Kumar, K. and Magnus, J.R. (2013) A characterization of Bayesian robustness for a normal location parameter. *Sankhya (Series B)* 75: 216–237.
Leeb, H. and Pötscher, B.M. (2003) The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19: 100–142.
Leeb, H. and Pötscher, B.M. (2005) Model selection and inference: Facts and fiction. *Econometric Theory* 21: 21–59.
Leeb, H. and Pötscher, B.M. (2006) Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34: 2554–2591.
Leeb, H. and Pötscher, B.M. (2008) Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* 24: 338–376.
Lenkoski, A., Eicher, T.S. and Raftery, A.E. (2013) Two-stage Bayesian model averaging in endogenous variable models. *Econometric Reviews* 33: 122–151.
Leung, G. and Barron, A.R. (2006) Information theory and mixing least-squares regressions. *IEEE Transactions and Information Theory* 52: 3396–3410.
Lewis, S.M. and Raftery, A.E. (1997) Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association* 92: 648–655.
Ley, E. and Steel, M.F.J. (2007) Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics* 29: 476–493.
Ley, E. and Steel, M.F.J. (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* 24: 651–674.
Ley, E. and Steel, M.F.J. (2012) Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* 171: 251–266.
Liang, H., Zou, G., Wan, A.T.K. and Zhang, X. (2011) Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association* 106: 1053–1066.
Liski, A., Liski, E.P., Sund, R. and Juntunen, M. (2010) A comparison of WALS estimation with pretest and model selection alternatives with an application to costs of hip fracture treatments. In K. Yamanishi, I. Kontoyiannis, E.P. Liski, P. Myllymäki, J. Rissanen and I. Tabus (eds), *Proceedings of the Third Workshop in Information Theoretic Methods in Science and Engineering (WITMSE)* (pp. 66–71). Finland: TICSP Series.
Liu, C.A. (2014) Distribution theory of the least squares averaging estimator. *Journal of Econometrics* doi:10.1016/j.jeconom.2014.07.002.
Liu, Q. and Okui, R. (2013) Heteroskedasticity-robust C_p model averaging. *Econometrics Journal* 16: 463–472.
Liu, Q., Okui, R. and Yoshimura, A. (2013) Generalized least squares model averaging. Discussion Paper 855, Kyoto Institute of Economic Research.
Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89: 1535–1546.
Magnus, J.R. (1999) The traditional pretest estimator. *Theory of Probability and Its Applications* 44: 293–308.
Magnus, J.R. (2002) Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5: 225–236.
Magnus, J.R. and Durbin, J. (1999) Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67: 639–643.
Magnus, J.R. and Vasnev, A. (2008) Using macro data to obtain better micro forecasts. *Econometric Theory* 24: 553–579.
Magnus, J.R. and Wang, W. (2014) Concept-based Bayesian model averaging and growth empirics. *Oxford Bulletin of Economics and Statistics* doi:10.1111/obes.12068.
Magnus, J.R., Powell, O. and Prüfer, P. (2010) A comparison of two averaging techniques with an application to growth empirics. *Journal of Econometrics* 154: 139–153.

Magnus, J.R., Wan, A.T.K. and Zhang, X. (2011) Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics & Data Analysis* 55: 1331–1341.
Magnus, J.R., Wang, W. and Zhang, X. (2014) Weighted-average least squares prediction. *Econometric Reviews* 33: to appear.
McCormick, T.H., Raftery, A.E., Madigan, D. and Burd, R.S. (2012) Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics* 68: 23–30.
Moral-Benito, E. (2012) Determinants of economic growth: A Bayesian panel data approach. *Review of Economics and Statistics* 94: 566–579.
Moral-Benito, E. (2013) Model averaging in economics: An overview. *Journal of Economic Surveys* doi:10.1111/joes.12044.
Osterloh, S. (2012) Words speak louder than actions: The impact of politics on economic performance. *Journal of Comparative Economics* 40: 318–336.
Poghosyan, K. and Magnus, J.R. (2012) WALS estimation and forecasting in factor-based dynamic models with an application to Armenia. *International Econometric Review* 4: 40–58.
Pötscher, B.M. (2006) The distribution of model averaging estimators and an impossibility result regarding its estimation. *IMS Lecture Notes — Monograph Series: Time Series and Related Topics* 52: 113–129.
Raftery, A.E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* 83: 251–266.
Sala-i-Martin, X., Doppelhofer, G. and Miller, R.I. (2004) Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94: 813–835.
Schomaker, M., Wan, A.T.K. and Heumann, C. (2010) Frequentist model averaging with missing observations. *Computational Statistics & Data Analysis* 54: 3336–3347.
Seya, H. and Tsutsumi, M. (2012) Application of model averaging techniques to spatial hedonic land price models. In S. A. Mendez and A. M. Vega (eds), *Econometrics: New Research* (pp. 63–88). London: Nova Science Publisher.
Seya, H., Tsutsumi, M. and Yamagata, Y. (2014) Weighted-average least squares applied to spatial econometric models: A Monte Carlo investigation. *Geographical Analysis* 46: 126–147.
Thomson, M. and Schmidt, P. (1982) A note on the comparison of the mean square error of inequality constrained least squares and other related estimators. *Review of Economics and Statistics* 64: 174–176.
Tierney, L. and Kadane, J.B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 81: 82–86.
Wan, A.T.K. and Zhang, X. (2009) On the use of model averaging in tourism research. *Annals of Tourism Research* 36: 525–532.
Wan, A.T.K., Zhang, X. and Zou, G. (2010) Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156: 277–283.
Wan, A.T.K., Zhang, X. and Wang, S. (2014) Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting* 30: 118–128.
Wang, H., Zhang, X. and Zou, G. (2009) Frequentist model averaging estimation: A review. *Journal of System Science and Complexity* 22: 732–748.
Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96: 574–588.
Yang, Y. (2003) Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* 13: 783–809.
Yang, Y. (2004) Combining forecasting procedures: Some theoretical results. *Econometric Theory* 20: 176–222.
Yuan, Z. and Yang, Y. (2005) Combining linear regression models: When and how? *Journal of the American Statistical Association* 100: 1202–1214.
Yüksel, G., Billor, N. and Ünal, D. (2010) Shrinkage pre-test estimator of the univariate normal mean. *Pakistan Journal of Statistics* 26: 461–477.
Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner (eds), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (pp. 233–243). Amsterdam: North-Holland.
Zhang, X. and Liang, H. (2011) Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics* 39: 174–200.

Zhang, X., Lu, Z. and Zou, G. (2013a) Adaptively combined forecasting for discrete response time series. *Journal of Econometrics* 176: 80–91.

Zhang, X., Wan, A.T.K. and Zhou, S.Z. (2012) Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business & Economic Statistics* 30: 132–142.

Zhang, X., Wan, A.T.K. and Zou, G. (2013b) Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics* 174: 82–94.

Zhang, X., Zou, G. and Liang, H. (2014) Model averaging and weight choice in linear mixed-effects models. *Biometrika* 101: 205–218.

Zou, H. and Yang, Y. (2004) Combining time series models for forecasting. *International Journal of Forecasting* 20: 69–84.

Zou, G., Wan, A.T.K., Wu, X. and Chen, T. (2007) Estimation of regression coefficients of interest when other regression coefficients are of no interest: The case of non-normal errors. *Statistics & Probability Letters* 77: 803–810.