Computational Statistics and Data Analysis 55 (2011) 1331–1341

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market

Jan R. Magnus [a], Alan T.K. Wan [b,*], Xinyu Zhang [c]

[a] Department of Econometrics & Operations Research, Tilburg University, Netherlands
[b] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong
[c] Academy of Mathematics & Systems Science, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 17 January 2010
Received in revised form 22 September 2010
Accepted 22 September 2010
Available online 29 September 2010

Keywords:
Model averaging
Bayesian analysis
Monte Carlo
Housing demand

ABSTRACT

The recently proposed 'weighted average least squares' (WALS) estimator is a Bayesian combination of frequentist estimators. It has been shown that the WALS estimator possesses major advantages over standard Bayesian model averaging (BMA) estimators: the WALS estimator has bounded risk, allows a coherent treatment of ignorance and its computational effort is negligible. However, the sampling properties of the WALS estimator as compared to BMA estimators are heretofore unexamined. The WALS theory is further extended to allow for nonspherical disturbances, and the estimator is illustrated with data from the Hong Kong real estate market. Monte Carlo evidence shows that the WALS estimator performs significantly better than standard BMA and pretest alternatives.

1. Introduction

In standard econometric practice, regression diagnostics such as the t-ratio are used to select a model, after which the unknown parameters are estimated in the selected model. These estimates are then presented as if they were unbiased and the standard errors as if they represented the true standard deviations. This is clearly wrong, even in the unlikely event that the selected model happens to be the data-generating process.

The traditional role of the t-ratio in model selection is itself suspect. The t-ratio was developed in the context of hypothesis testing, not in the context of model selection. While it may make good sense to reject a null hypothesis only when 95% of the evidence is in favor of the alternative, it does not make nearly as good sense to only prefer one model over another when 95% of the evidence points that way.

Even when we accept the t-ratio as a model selection tool, there is the problem of pretesting. We are ignoring the fact that model selection and estimation are a combined effort and that the model selection part influences the properties of our estimators. These properties depend not only on the stochastic nature of our framework, but also on the way the model has been selected. If we ignore the model selection part, then our reported properties are conditional rather than unconditional. Focusing on the general-to-specific and specific-to-general model selection procedures, Danilov and Magnus (2004a) derived the unconditional moments of the least squares pretest estimators, and Danilov and Magnus (2004b) obtained the corresponding forecast moments. In a series of papers Leeb and Pötscher (2003, 2005, 2006, 2008) also studied conditional distributions of estimators subsequent to a range of model selection strategies. All these studies conclude that the problems associated with ignoring model selection can be very serious. For example, Danilov and Magnus (2004a)

* Corresponding author. Tel.: +852 27887146; fax: +852 27888560.
E-mail addresses: magnus@uvt.nl (J.R. Magnus), msawan@cityu.edu.hk (A.T.K. Wan), xinyu@amss.ac.cn (X. Zhang).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.09.023

showed that the mean squared error associated with the specific-to-general pretest procedure is unbounded, and that the errors in not reporting the correct moments can be large.

But even if we could fully understand the properties of the pretest estimator and take them into account, the problem remains that the pretest estimator itself is a poor estimator. This is primarily because the pretest estimator is 'kinked', which implies that it is inadmissible (Magnus, 2002, Theorem A.6). In addition to this theoretical problem, the pretest estimator is also uncomfortable intuitively. We may be used to taking one estimator when $t = 1.95$ and another when $t = 1.97$, but it remains difficult to justify. Would it not be more comfortable to abandon this 'kinked' estimator and use a 'continuous' estimator instead? A 'continuous' estimate of a specific parameter would consider the estimates from all models in a certain model space and then weigh them according to their statistical strengths, possibly in addition to prior information. This is the idea underlying model averaging.

The aim of model averaging is not to find the best possible model but rather to find the best estimates of the parameters of interest. Instead of selecting a single 'winning' model, we use information from all models, but of course some models are more reliable than others, based on data and priors. Model uncertainty is then incorporated in the properties of our estimators in a natural way. Bayesian model averaging (BMA) offers a natural framework (Leamer, 1978) and has been used widely, especially in growth econometrics. Raftery et al. (1997) and Hoeting et al. (1999) provide useful literature summaries of BMA. Recent applications of BMA can be found in Peña and Redondas (2006) and Ouysse and Kohn (2010). There has also been a rising interest in model averaging from a frequentist perspective and several frequentist model average estimators have been proposed. For example, Buckland et al. (1997) suggest mixing models based on the AIC or BIC scores of the competing models; Yang (2001, 2003) propose a frequentist-based adaptive regression mixing method; Hjort and Claeskens (2003) provide a likelihood-based local misspecification framework for analyzing the asymptotic properties of model average estimators; Hansen (2007, 2008) and Wan et al. (2010) suggest a Mallows criterion for selecting weights in a model average estimator; and Schomaker et al. (2010) develop frequentist model averaging schemes in the face of missing observations. The recent monograph by Claeskens and Hjort (2008) covers much of the progress that has been made in this direction. In addition, subset selection has been proposed in the frequentist model averaging literature, in order to narrow down the list of candidate models for averaging and thus save computing time; see, for example, Yuan and Yang (2005), Claeskens et al. (2006) and Buchholz et al. (2008).

Recently, Magnus et al. (2010), hereafter MPP, proposed a weighted average least squares (WALS) estimator based on the equivalence theorem developed in Magnus and Durbin (1999) and Danilov and Magnus (2004a). The WALS estimator is a Bayesian combination of frequentist estimators, and it possesses some important advantages over standard BMA techniques. First, in contrast to standard BMA estimators that adopt normal priors leading to unbounded risk, WALS uses priors from the Laplace distribution and thus generates bounded risk. Second, the use of Laplace priors implies a coherent treatment of ignorance. Third, WALS requires a trivial computational effort because computing time is linear in the number of regressors rather than exponential as in BMA or standard frequentist model averaging.

MPP discuss WALS estimation in the context of growth models, where there is a large number of potentially relevant explanatory variables, while Wan and Zhang (2009) apply WALS in a tourism study. In both cases the disturbances are assumed to be identically and independently distributed. This assumption is not realistic in most applications, and one purpose of the current paper is to develop WALS in the more general framework of nonspherical disturbances. The main purpose of the paper is to provide Monte Carlo evidence on the performance of WALS in a realistic set-up. In MPP it is shown that WALS works well, has intuitive appeal, is easy to compute, and produces realistic results, but we do not know yet whether the estimates are close to the true parameter values, neither do we know whether the standard errors capture the combined model selection and estimation errors adequately. This can only be found out through simulations and the current study attempts to provide a representative subset of such simulations.

As our empirical framework we choose a hedonic housing price model with data from Hong Kong, one of the world's most buoyant real estate markets. There typically exists a wide range of model specifications in hedonic housing price modeling, and this makes model uncertainty an important and challenging issue. We explore the root mean squared error (RMSE) of the WALS estimator and contrast this with the RMSE of the BMA estimator and a *stepwisefit* pretest estimator. We conclude that the WALS estimator has smaller RMSE than BMA and other model selection alternatives over a large portion of the parameter space, and is therefore a potentially useful estimator, not only when the number of regressors is large, but in any linear regression problem.

Our theoretical framework is the linear regression model

$$y = X_1\beta_1 + X_2\beta_2 + u = X\beta + u, \tag{1}$$

where y $(n \times 1)$ is the vector of observations on the dependent variable, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are matrices of nonrandom regressors, u is a random vector of unobservable disturbances, and β_1 and β_2 are unknown parameter vectors. The columns of X_1 are called focus regressors (those we want in the model on theoretical or other grounds), while the columns of X_2 are called auxiliary regressors (those we are less certain of). We assume that $k_1 \geq 0,\ k_2 \geq 0,\ 1 \leq k := k_1 + k_2 \leq n - 1$, and that $X := (X_1 : X_2)$ has full column-rank. Model selection takes place over the auxiliary regressors only. Since each of the k_2 auxiliary regressors can either be included or not, we have 2^{k_2} models to consider. In contrast to common practice in the model averaging literature we shall not assume that the disturbances are identically and independently distributed. Instead we assume that

$$u \sim \mathrm{N}(0, \Omega(\theta)), \tag{2}$$

where $\Omega(\theta)$ is a positive definite $n \times n$ matrix whose elements are functions of an m-dimensional unknown parameter vector $\theta = (\theta_1, \ldots, \theta_m)'$. The situation where the variance matrix Ω is singular or ill-conditioned and the dependent variable cannot be easily transformed (reduced) so that the variance becomes nonsingular, is not discussed in this paper.

The plan of this paper is as follows. In Section 2 we extend the results in MPP to nonspherical disturbances, generalize the equivalence theorem, and explain the Laplace priors. Section 3 discusses the background of the Hong Kong real estate market and the data used in the empirical study and presents the parameter estimates based on maximum likelihood and WALS procedures. Section 4 describes the design of the Monte Carlo study focusing on the finite-sample performance of WALS, BMA, and pretest estimators, and reports the results of this comparison. Section 5 concludes.

2. The WALS estimator

We assume (at first) that Ω is known. It is convenient to semi-orthogonalize the regression model as follows. Let

$$M_1^* := \Omega^{-1} - \Omega^{-1}X_1(X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1},$$

where we notice that while M_1^* is not idempotent, the matrix $\Omega^{1/2}M_1^*\Omega^{1/2}$ is idempotent. Let P be an orthogonal matrix and Λ a diagonal matrix with positive diagonal elements such that $P'X_2'M_1^*X_2P = \Lambda$. Next define the transformed auxiliary regressors and the transformed auxiliary parameters as

$$X_2^* := X_2P\Lambda^{-1/2} \quad \text{and} \quad \beta_2^* := \Lambda^{1/2}P'\beta_2,$$

respectively. Then $X_2^*\beta_2^* = X_2\beta_2$, so that we can write model (1) equivalently as $y = X_1\beta_1 + X_2^*\beta_2^* + u$. The result of this transformation is that the new design matrix $(X_1 : X_2^*)$ is 'semi-orthogonal' in the sense that $X_2^{*\prime}M_1^*X_2^* = I_{k_2}$ and this has important advantages which will become clear shortly. Our strategy will be to estimate (β_1, β_2^*) rather than (β_1, β_2). However, the estimator of β_2 and its distribution can always be recovered through $\beta_2 = P\Lambda^{-1/2}\beta_2^*$.

Each component of β_2^* can either be included or not included in the model and this gives rise to 2^{k_2} models. A specific model is identified through a $k_2 \times (k_2 - k_{2i})$ selection matrix S_i of full column-rank, where $0 \le k_{2i} \le k_2$, so that $S_i' = (I_{k_2-k_{2i}} : 0)$ or a column permutation thereof. Our first interest is in the generalized least squares (GLS) estimator of (β_1, β_2^*) in the ith model, that is, the GLS estimator of (β_1, β_2^*) under the restriction $S_i'\beta_2^* = 0$.

Let $\mathcal{M}_i$ represent model (1) under the restriction $S_i'\beta_2^* = 0$, and let $\hat\beta_{1(i)}$ and $\hat\beta_{2(i)}^*$ denote the GLS estimators of β_1 and β_2^* under $\mathcal{M}_i$. Extending Danilov and Magnus (2004a, Lemmas A1 and A2), the restricted GLS estimators of β_1 and β_2^* may be written as

$$\hat\beta_{1(i)} = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y - Q^*W_ib_2^*, \qquad \hat\beta_{2(i)}^* = W_ib_2^*, \tag{3}$$

respectively, where $b_2^* := X_2^{*\prime}M_1^*y$ is the GLS estimator of β_2^* in the unrestricted model, $Q^* := (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}X_2^*$ and $W_i := I_{k_2} - S_iS_i'$ is a *diagonal* $k_2 \times k_2$ matrix with k_{2i} ones and $(k_2 - k_{2i})$ zeros on the diagonal; the jth diagonal element of W_i equals zero if β_{2j}^* (the jth component of β_2^*) is restricted to zero, and it equals one otherwise. Note that if $k_{2i} = k_2$ then $W_i = I_{k_2}$. The diagonality of W_i, which is a direct consequence of the semi-orthogonal transformation, is an essential ingredient in our analysis.

The distributions of $\hat\beta_{1(i)}$ and $\hat\beta_{2(i)}^*$ are given by

$$\hat\beta_{1(i)} \sim \mathrm{N}_{k_1}(\beta_1 + Q^*S_iS_i'\beta_2^*,\ (X_1'\Omega^{-1}X_1)^{-1} + Q^*W_iQ^{*\prime}),$$

and

$$\hat\beta_{2(i)}^* \sim \mathrm{N}_{k_2}(W_i\beta_2^*,\ W_i),$$

respectively, and the two estimators are correlated with $\mathrm{cov}(\hat\beta_{1(i)}, \hat\beta_{2(i)}^*) = -Q^*W_i$. The residual vector takes the form

$$e_i := y - X_1\hat\beta_{1(i)} - X_2^*\hat\beta_{2(i)}^* = \Omega D_i^*y,$$

where $D_i^* := M_1^* - M_1^*X_2^*W_iX_2^{*\prime}M_1^*$ and $\Omega^{1/2}D_i^*\Omega^{1/2}$ is a symmetric idempotent matrix of rank $n - k_1 - k_{2i}$. It follows that: (1) all models that include the jth column of X_2^* as a regressor have the same estimator of β_{2j}^*, namely b_{2j}^*, irrespective of which other columns of X_2^* are included; (2) the estimators $b_{21}^*, b_{22}^*, \ldots, b_{2k_2}^*$ are independent; and (3) the residuals in the ith model $\mathcal{M}_i$ depend on y only through M_1^*y.

We now define the WALS estimator of β_1 as

$$b_1 = \sum_{i=1}^{2^{k_2}} \lambda_i\hat\beta_{1(i)}, \tag{4}$$

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

$$\lambda_i \ge 0, \qquad \sum_i \lambda_i = 1, \qquad \lambda_i = \lambda_i(M_1^*y). \tag{5}$$

The WALS estimator of β_1 can then be written as

$$b_1 = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y - Q^*Wb_2^*,$$

where $W := \sum_i \lambda_i W_i$. While the W_i's are nonrandom diagonal matrices, the matrix W is random (because it depends on λ_i's) but still diagonal.

By writing $\tilde{X} = \Omega^{-1/2}X$ and $\tilde{y} = \Omega^{-1/2}y$, we obtain the following theorem which generalizes the results of Magnus and Durbin (1999), Danilov and Magnus (2004a, Theorem 1), and MPP (Section 3.3) to the case of nonspherical disturbances.

Theorem 1 (*Equivalence Theorem for Nonspherical Disturbances*)**.** *If the weights λ_i satisfy condition (5), then*

$$\mathrm{E}(b_1) = \beta_1 - Q^*\,\mathrm{E}(Wb_2^* - \beta_2^*), \qquad \mathrm{var}(b_1) = (X_1'\Omega^{-1}X_1)^{-1} + Q^*\,\mathrm{var}(Wb_2^*)Q^{*\prime},$$

and

$$\mathrm{MSE}(b_1) = (X_1'\Omega^{-1}X_1)^{-1} + Q^*\,\mathrm{MSE}(Wb_2^*)Q^{*\prime}.$$

The equivalence theorem implies that the WALS estimator b_1 will be a 'good' estimator of β_1 in the mean squared error sense if and only if Wb_2^* is a 'good' estimator of β_2^*. That is, if we can find λ_i's such that Wb_2^* is an 'optimal' estimator of β_2^*, then *the same* λ_i's will provide an 'optimal' estimator of β_1. We do not know W, but we do know (because of the semi-orthogonal transformation) that W is diagonal, say $W = \mathrm{diag}(w_1, \ldots, w_{k_2})$, and that the k_2 components b_{2j}^* of b_2^* are independent with $\mathrm{var}(b_{2j}^*) = 1$. Therefore, if we choose w_j to be a function of b_{2j}^* only, then the components $w_j b_{2j}^*$ of Wb_2^* will also be independent, and our k_2-dimensional problem reduces to k_2 one-dimensional problems. This one-dimensional problem is simply to estimate β_{2j}^* using information from the data that $b_{2j}^* \sim \mathrm{N}(\beta_{2j}^*, 1)$, possibly together with prior information on β_{2j}^*.

The question then is how to estimate β_{2j}^*. This seemingly trivial question was addressed in Magnus (2002) and in MPP. The 'usual' estimator $t(b_{2j}^*) = b_{2j}^*$ is unbiased, admissible, and minimax. It has good properties when $|\beta_{2j}^*|$ is large but not when β_{2j}^* is close to zero. The 'silly' estimator $t(b_{2j}^*) = 0$ has excellent properties when β_{2j}^* is close to zero, but its risk increases without bound when $|\beta_{2j}^*|$ becomes large. The pretest estimator, $t(b_{2j}^*) = b_{2j}^*$ if $|b_{2j}^*| > c$ and $t(b_{2j}^*) = 0$ if $|b_{2j}^*| \leq c$ for some $c \geq 0$, has bounded risk, but it has a discontinuity at c and is therefore inadmissible.

We can also consider Bayesian estimators. If we combine the normal likelihood $b_{2j}^* \mid \beta_{2j}^* \sim \mathrm{N}(\beta_{2j}^*, 1)$ with a normal prior $\pi(\beta_{2j}^*) \sim \mathrm{N}(0, 1/c)$, we obtain $t(b_{2j}^*) = \mathrm{E}(\beta_{2j}^*|b_{2j}^*) = b_{2j}^*/(1+c)$, which is essentially the BMA estimator. The risk of this estimator is unbounded. We propose a Bayesian estimator whose risk remains bounded. This is achieved by combining the normal likelihood with a Laplace prior $\pi(\beta_{2j}^*) = (c/2)\exp(-c|\beta_{2j}^*|)$ rather than with a normal prior. The resulting 'Laplace' estimator is admissible, has bounded risk, has good properties around $|\beta_{2j}^*| = 1$, and is near-optimal in terms of minimax regret; see Magnus (2002). The hyperparameter c is chosen as $c = \log 2$, because this implies

$$\Pr(\beta_{2j}^* > 0) = \Pr(\beta_{2j}^* < 0) \quad \text{and} \quad \Pr(|\beta_{2j}^*| > 1) = \Pr(|\beta_{2j}^*| < 1).$$

What this means is that we assume a priori ignorance about whether β_{2j}^* is positive or negative, and also about whether $|\beta_{2j}^*|$ is larger or smaller than one. These seem natural properties for a prior in our context, because we do not know a priori whether the β_2^* coefficients are positive or negative, and we also do not know whether adding a specific column of X_2^* to the model will increase or decrease the mean squared error of the estimators of the focus parameters; see Theorem 1 of Magnus and Durbin (1999). Such a prior thus captures prior ignorance in a natural way. The mean and variance of the posterior distribution of $\beta_{2j}^*|b_{2j}^*$ are given in Theorem 1 of MPP.

Let $\tilde{b}_{2j}^* := \mathrm{E}(\beta_{2j}^*|b_{2j}^*)$ denote the Laplace estimator of β_{2j}^* and $\xi_j^2 := \mathrm{var}(\beta_{2j}^*|b_{2j}^*)$ the associated variance from the posterior distribution. Define $\tilde{b}_2^* := (\tilde{b}_{21}^*, \ldots, \tilde{b}_{2k_2}^*)'$ and $\Xi := \mathrm{diag}(\xi_1^2, \ldots, \xi_{k2}^2)$. The equivalence theorem gives us the corresponding WALS estimator of β_1 and its variance as

$$b_1 = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}y - Q^*\tilde{b}_2^* = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}(y - X_2^*\tilde{b}_2^*),$$

and

$$\mathrm{var}(b_1) = (X_1'\Omega^{-1}X_1)^{-1} + Q^*\,\mathrm{var}(\tilde{b}_2^*)Q^{*\prime},$$

respectively. Furthermore, we obtain the estimates of β_2 through the relation $\beta_2 = P\Lambda^{-1/2}\beta_2^*$. This gives

$$b_1 = (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}(y - X_2b_2), \qquad b_2 = P\Lambda^{-1/2}\tilde{b}_2^*,$$

using the fact that $X_2^*\tilde{b}_2^* = X_2b_2$. Letting $Q := (X_1'\Omega^{-1}X_1)^{-1}X_1'\Omega^{-1}X_2$, we obtain the variance of b_1 and b_2 as

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

and

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

and the covariance as $\mathrm{cov}(b_1, b_2) = -Q\,\mathrm{var}(b_2)$.

So far we have assumed that Ω is known, whereas in practice Ω is of course unknown. One may estimate θ based on the unrestricted model by minimizing

$$\varphi(\theta) := \log|\Omega| + y'(\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1})y$$

with respect to θ. This leads to the maximum likelihood estimator $\hat{\theta}$ of θ, through which we also obtain an estimator $\hat{\Omega} = \Omega(\hat{\theta})$ of Ω. Note that the gradient of φ is the $m \times 1$ vector whose ith component is given by

$$\frac{\partial\varphi(\theta)}{\partial\theta_i} = \operatorname{tr}\left(\Omega^{-1}\frac{\partial\Omega}{\partial\theta_i}\right) - (M^*y)'\frac{\partial\Omega}{\partial\theta_i}(M^*y),$$

where

$$M^* = M_1^*(\Omega - X_2^*X_2^{*\prime})M_1^*.$$

Therefore, $\hat{\theta}$ depends on y only through M_1^*y and the same holds for $\hat{\Omega}$.

In summary, we have developed an estimator (the WALS estimator) whose properties incorporate the fact that model selection and estimation are a combined effort and that therefore the noise generated by model selection should be explicitly reflected in the bias and variance of the estimator. The WALS estimator allows for nonspherical disturbances, and uses a prior on the auxiliary parameters (the Laplace prior) which captures ignorance in a natural way. The computational effort is negligible even when the number of regressors is large. To emphasize the last point, empirical research on the determinants of economic growth has identified numerous variables as being potentially relevant to productivity growth (Durlauf et al. (2005) list 145 potential right-hand side variables for growth regressions), and MPP provide WALS estimates for a system with 67 regressors.

The computational code requires a number of simple steps which are, in essence, given in Section 3.5 of MPP, except that we use M_1^* instead of M_1. The code is available from http://center.uvt.nl/staff/magnus/wals, the WALS' website, which also provides further examples and links.

It may be worth mentioning that the idea underlying WALS is closely related to multiple shrinkage estimation (George, 1986), which combines shrinkage estimators for different shrinking targets by a weight function, and thus permits the parameter estimates to be shrunk to more than one subspace simultaneously. A subsequent study by Adkins (1996) proposes a weight choice based on Stein's unbiased estimator of the risk.

3. The Hong Kong housing market

Together with London, Moscow, New York, Tokyo, and Paris, Hong Kong is one of the most expensive cities in the world in terms of housing cost (Global Property Guide, 2008), and several studies have recently analyzed the Hong Kong property market; see Bao and Wan (2004, 2007), Wong et al. (2005) and Hui and Yue (2006). Since the 1980s, the Hong Kong residential property market has been dominated by private housing estates characterized by clusters of high-rise buildings with 30–40 floors, each floor containing four to six apartments (more precisely, condominiums). These estates often have their own shopping centers, and many of the newer estates are equipped with leisure facilities such as swimming pools, gymnasiums, and tennis courts. Apartments in the same estate tend to be very similar. It is not unusual for a housing estate with more than one thousand apartments to offer only five or six different floor plans, and many apartments in the same estate are therefore identical in terms of living area, number of bedrooms, and kitchen size. Apartments in the same housing estate also have similar access to public transport and other facilities.

Even though apartments are often similar, there is price heterogeneity within the same estate, and this is largely determined by floor plan, floor level, and direction. The floor plan affects the size of the apartment; the floor level determines not only the view, but also the amount of pollution and noise resulting from traffic; and the direction is important for the view.

Our data are based on sales at the housing estate 'South Horizon' located in the South of Hong Kong near 'Ocean Park' and the industrial district of Aberdeen. The data, together with characteristics of the transacted apartments, are extracted from the Centa-City Index database provided by Centaline Property Agency Ltd. The Centa-City Index is a publicly available index for tracking weekly apartment price movements in Hong Kong. South Horizon is made up of 9812 apartments in 34 blocks, each with 25–42 floors. The floor areas of the apartments range from 632 to 1121 square feet. The majority of the apartments have 2–3 bedrooms and 1–2 bathrooms. Apartments between blocks may vary in age as the estate construction was completed in four phases between 1991 and 1995. Our observation period is January 2004 to October 2007, and during this period Centaline Property Agency Ltd. recorded $n = 560$ transactions, that is, about twelve transactions per month.

The dependent variable in our analysis is the natural logarithm of the sales price per square foot (*LPRICE*). There are twelve regressors, including the constant term, as shown in Table 1. The regressors contain information about the size (*LAREA*) and the floor level (*LFLOOR*), and much information about the view, which is deemed very important in the Hong Kong housing market. The apartment is more attractive if it has a view of one of several gardens located inside the estate (*GARV*), and less attractive if factories or industrial plants of nearby Aberdeen are visible from the apartment (*INDV*). A view of the bay Shek Pai Wan or of East Lamma Channel is also attractive, and we distinguish between a full sea view (*SEAVF*), a semi-sea view (*SEAVS*), and a minor sea view (*SEAVM*). Some apartments have a view of Brick Hill in the East, where Ocean Park is located

Table 1
Explanation of the regressors.

Regressors	Explanation
INTERCEPT	Constant term
LAREA	Size of apartment in square feet (natural logarithm)
LFLOOR	Floor level of apartment (natural logarithm)
GARV	1 if garden view; 0 otherwise
INDV	1 if industry view; 0 otherwise
SEAVF	1 if full sea view; 0 otherwise
SEAVS	1 if semi-sea view; 0 otherwise
SEAVM	1 if minor sea view; 0 otherwise
MONV	1 if mountain view; 0 otherwise
STRI	1 if internal street view; 0 otherwise
STRN	1 if no street view; 0 otherwise
UNLUCK	1 if located on floors 4, 14, 24, 34 or in block 4; 0 otherwise

Table 2
Estimation results: ML, WALS, Pretest (PT), and BMA.

Regressors	ML Est. (std)	WALS Est. (std)	PT Est. (std)	BMA Est. (std)
Focus regressors				
INTERCEPT	3.73 (0.39)	3.83 (0.36)	3.89 (0.33)	3.85 (0.28)
LAREA	0.66 (0.06)	0.65 (0.05)	0.64 (0.05)	0.65 (0.04)
LFLOOR	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
GARV	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
INDV	−0.07 (0.03)	−0.07 (0.03)	−0.07 (0.03)	−0.07 (0.02)
SEAVF	0.05 (0.02)	0.05 (0.02)	0.05 (0.02)	0.04 (0.02)
Auxiliary regressors				
SEAVS	−0.04 (0.03)	−0.03 (0.02)	−0.05 (0.03)	−0.04 (0.03)
SEAVM	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.01 (0.02)
MONV	−0.01 (0.02)	−0.01 (0.02)	–	0.00 (0.01)
STRI	0.02 (0.15)	0.01 (0.11)	–	0.00 (0.01)
STRN	0.03 (0.12)	0.02 (0.09)	0.02 (0.07)	0.00 (0.01)
UNLUCK	−0.02 (0.02)	−0.01 (0.02)	–	0.00 (0.01)
Heteroskedasticity				
θ	−4.32 (0.05)	−4.32 (0.05)	−4.32 (0.05)	−4.32 (0.05)

(*MONV*). Most people do not find this view particularly pleasant, because Brick Hill is quite a distance from South Horizon and the view of it also includes the view of some government housing estates located midway between South Horizon and Ocean Park. The apartment may look onto streets within the estate (*STRI*) or it may not look onto any streets (*STRN*); the latter is more attractive because a street view implies more noise. Finally, the variable *UNLUCK* is included for block or floor levels associated with the number four because traditional Cantonese culture links the number four to bad luck as four in Cantonese is pronounced as 'Si' which sounds like 'death'.

The model is first estimated by ordinary least squares (OLS) using all $n = 560$ observations. A priori we would expect heteroskedasticity since the data involve a heterogeneity of apartments. On plotting the OLS squared residuals against the regressor *STRN*, we find that the residuals are more spread out for $STRN = 1$ than for $STRN = 0$. For the other regressors, the residuals appear to scatter randomly without any systematic pattern. The Breusch and Pagan (1979) test decisively rejects homoskedasticity when *STRN* is taken as the regressor that causes the heteroskedasticity. Therefore we assume multiplicative heteroskedasticity of the form

$$\sigma_i^2 = \exp(\theta\, STRN_i) \quad (i = 1, \ldots, n).$$

We then re-estimate all parameters in the model by maximum likelihood (ML), and obtain the estimates and standard errors as reported in the panel labeled 'ML' of Table 2. (The distinction between focus and auxiliary regressors is not relevant here, but it is relevant for WALS, pretest, and BMA.) The estimates show that six of the twelve parameters are significant at the 5% level, namely *LAREA*, *LFLOOR*, *GARV*, *INDV*, *SEAVF*, and the intercept. The *SEAVM* parameter is significant at the 10% level, but all other parameters are not significant at any conventional level. The signs of the estimates are consistent with prior expectations. More space, a higher floor level, and a garden view are positive features for an apartment, while an industry or mountain view or an 'unlucky' floor number are negative features. A sea view is a positive feature, and within this class a full sea view is preferred over a semi-sea view, which in turn is preferred over a minor sea view. A view on an internal street and no street view are also positive features.

We now treat the six significant variables as focus regressors and the six insignificant variables as auxiliary regressors, and re-estimate the coefficients by WALS, again under heteroskedasticity. The resulting estimates and standard errors are reported in the panel labeled 'WALS' of Table 2. The WALS estimates have the same signs as and are close to the maximum likelihood estimates. In all cases, the standard errors of the estimates produced by WALS are smaller than those obtained by maximum likelihood. The estimate (and standard error) for θ is the same under WALS and ML, because WALS variance estimates are taken to be the ML estimates.

The pretest estimator is based on the *stepwisefit* routine in Matlab. This routine begins with a forward selection procedure based on an initial model, then employs a backward selection procedure to remove variables. The steps are repeated until no additions or deletions of variables are indicated. In implementing the routine, we treat the model that includes only the focus regressors as the initial model and let the routine select the auxiliary regressors according to statistical significance. We choose the significance level for adding a variable to be 0.05 and for removing a variable to be 0.10. The BMA estimator, generalized to allow for a distinction between focus and auxiliary regressors, was developed in MPP (Section 2) for the case where $u \sim \mathrm{N}(0, \sigma^2 I_n)$. Here we have $u \sim \mathrm{N}(0, \Omega)$, and hence, letting $\Omega = \sigma^2 V$, we need to estimate V by $\hat{V}$, and transform X and y by $\tilde{X} = \hat{V}^{-1/2}X$ and $\tilde{y} = \hat{V}^{-1/2}y$, respectively. We then use BMA on the transformed variables. The pretest and BMA estimates and standard errors reported in Table 2 show that the focus parameter estimates are again close to ML and WALS, but that the standard errors tend to be smaller for the pretest estimates and even smaller for BMA. The same is true for the auxiliary parameters, where in some cases (*STRI* and *STRN*) the difference in standard errors is large.

These results demonstrate that the WALS theory can be implemented without any problem and that the estimates and standard errors are comparable in size to the ML results. But the application does not prove that the WALS estimates and standard errors are more accurate than the ones currently in use. In particular, are the relatively low standard errors of the pretest and BMA estimates to be trusted or do they underestimate the 'true' variation in the parameter estimates? This question cannot be answered in an estimation exercise. We need Monte Carlo evidence where the estimates and their precisions are known.

4. Monte Carlo simulations

The purpose of the simulations is to investigate the performance of the WALS estimator versus Bayesian model averaging (BMA), pretest (PT), and maximum likelihood (ML) estimators within a realistic set-up. We use the ML estimator in the unrestricted model as our benchmark. The WALS estimator was described in Section 2; the BMA and the pretest estimator in Section 3.

The context of our experiment is provided by the hedonic housing price model and the Hong Kong data studied in the previous section. The six focus regressors X_1 and the six auxiliary regressors X_2 are thus known to us. In the simulations we know the values of β_1 and β_2, and we use the ML estimates as the true values of the parameters; see the left panel of Table 2. We also know the structure of Ω (heteroskedastic).

Following Adkins and Eells (1995), we do not generate the disturbances randomly; instead the disturbances are obtained by resampling the OLS residuals. The simulations are thus based on real rather than generated data which increases the realism of the experiment. The disturbances in each round of the experiment are obtained by randomly drawing 560 numbers with replacement from the OLS residuals. To preserve the assumed heteroskedastic structure, we apply resampling with the same pattern to the regressor *STRN*. Denote the vector of residuals and observations of *STRN* of the ℓth such sample as u_ℓ and $STRN_\ell$, respectively. The ℓth Monte Carlo sample is then generated by

$$y_\ell = X_1\beta_1 + X_{2\ell}\beta_2 + u_\ell,$$

where the dependent variable y_ℓ denotes the logarithm of the price (*LPRICE*) and $X_{2\ell}$ is the same as X_2 except that it contains $STRN_\ell$ instead of *STRN*. Because the simulations are based on resampling of the initial residuals rather than on generated values of the disturbances, we need not assume a value for θ.

In order to gain further insight we use different values of the β_2-vector by scaling it. More precisely, we replace β_2 by $\tau\beta_2$ where the scale factor τ is determined by

$$\phi = \tau^2\beta_2'X_2'M_1X_2\beta_2$$

and we let ϕ vary between 0 and 5. The interpretation of ϕ is that it approximates the theoretical F-ratio $\beta_2'X_2'M_1X_2\beta_2/(k_2\sigma^2)$.

The investigator knows which regressors are focus (X_1) and which are auxiliary (X_2) and he/she has data (y, X_1, X_2). The investigator also knows the *structure* of Ω (but not the value of θ), and uses the linear model

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

to estimate the β-parameters and their precisions. Each of the four estimation methods provides an algorithm for obtaining estimates and standard errors of the β-parameters. We concentrate on the $k_1 = 6$ focus parameters β_1 and thus obtain parameter estimates $\hat{\beta}_{1j}$ and standard errors $s_j := (\widehat{\mathrm{var}}(\hat{\beta}_{1j}))^{1/2}$ for $j = 1, \ldots, k_1$. Based on $R = 5000$ replications we approximate the distributions of $\hat{\beta}_{1j}$ and s_j from which we can compute moments of interest. The estimators are evaluated in terms of the root mean squared error (RMSE) of the parameter estimates as well as the RMSE of their standard errors. More precisely, letting $\hat{\beta}_{1j}^{(r)}$ and $s_j^{(r)}$ denote the estimates in the rth replication, we compute

$$\mathrm{BIAS}(\hat{\beta}_{1j}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{1j}^{(r)} - \beta_{1j}), \qquad \mathrm{RMSE}(\hat{\beta}_{1j}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{1j}^{(r)} - \beta_{1j})^2},$$

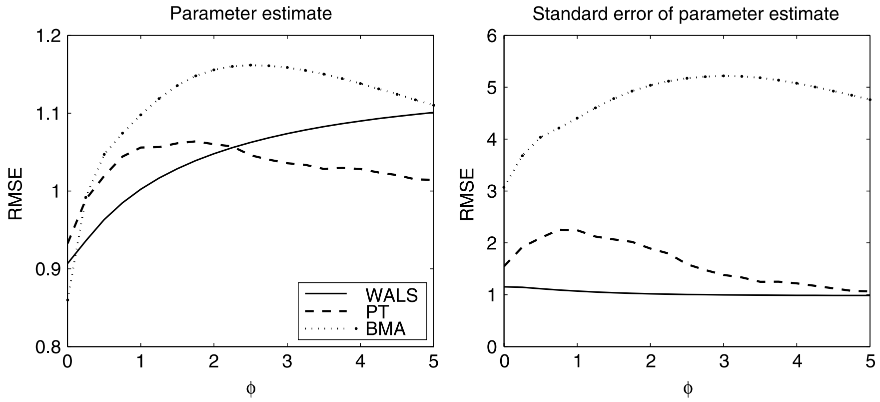

Fig. 1. RMSE comparisons for INTERCEPT.

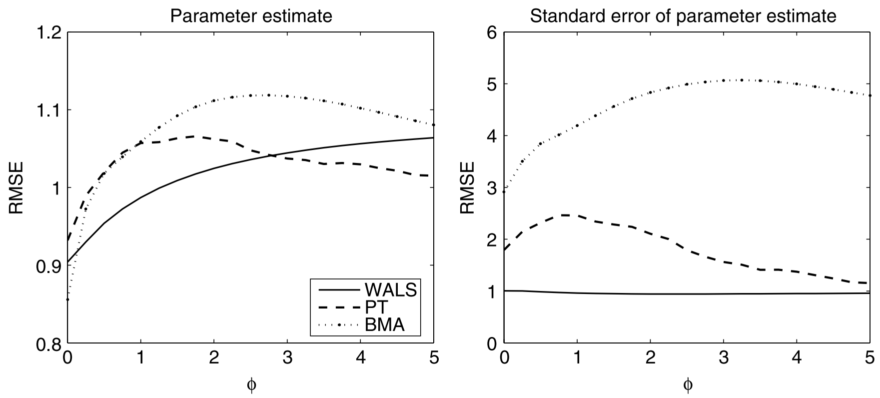

Fig. 2. RMSE comparisons for LAREA.

from which we find the standard error as

$$\mathrm{SE}(\hat{\beta}_{1j}) = \sqrt{\mathrm{RMSE}^2(\hat{\beta}_{1j}) - \mathrm{BIAS}^2(\hat{\beta}_{1j})}.$$

In this way we obtain the RMSE, BIAS, and SE of $\hat{\beta}_{1j}$ for each j.

We are not only interested in the distribution of $\hat{\beta}_1$ but also in its estimated precision. We want to know whether the reported precision is close to the actual precision of the estimator. Thus we compute

$$S_{1j} := \frac{1}{R}\sum_{r=1}^{R} s_j^{(r)}, \qquad S_{2j} := \frac{1}{R}\sum_{r=1}^{R} (s_j^{(r)})^2,$$

from which we obtain

$$\mathrm{SE}(s_j) = \sqrt{S_{2j} - S_{1j}^2}.$$

In order to find the bias and mean squared error of s_j we need to know the 'true' value of s_j. This value is approximated by $\mathrm{SE}(\hat{\beta}_{1j})$ obtained from the empirical distribution of the estimated β_{1j} from the Monte Carlo simulations. Thus we find

$$\mathrm{BIAS}(s_j) = S_{1j} - \mathrm{SE}(\hat{\beta}_{1j}), \qquad \mathrm{RMSE}(s_j) = \sqrt{\mathrm{SE}^2(s_j) + \mathrm{BIAS}^2(s_j)}.$$

Our results are summarized in Figs. 1–6, where each RMSE value is scaled by the corresponding RMSE of the ML estimator.

We first discuss the RMSEs of the parameter estimates (left panels). As shown in the figures, the comparisons are in favor of WALS. When estimating the parameters of *LFLOOR*, *GARV*, *INDV*, and *SEAVF*, the WALS estimator dominates the BMA and PT estimators almost uniformly; in all of these cases, the BMA estimator has a slight edge over WALS only when ϕ is near zero. When estimating the parameters of *INTERCEPT* and *LAREA*, the WALS estimator again yields smaller RMSE than the BMA estimator in nearly the entire region of the parameter space except when ϕ is near zero; the PT estimator, on the other hand, has better RMSE performance than both WALS and BMA estimators when ϕ is large.

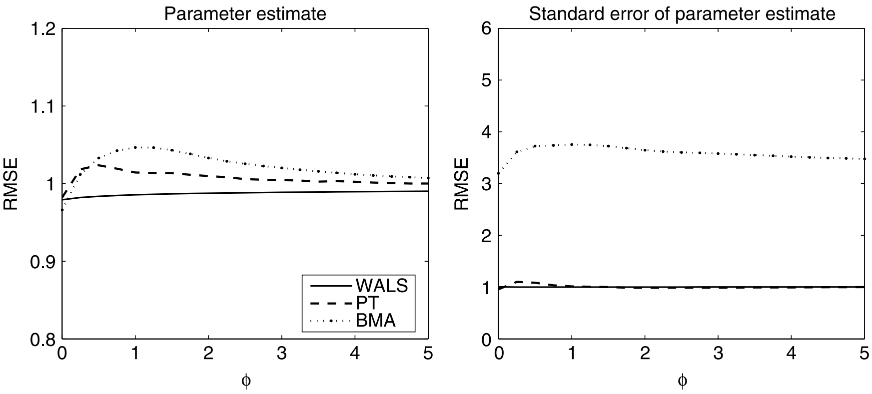

Fig. 3. RMSE comparisons for LFLOOR.

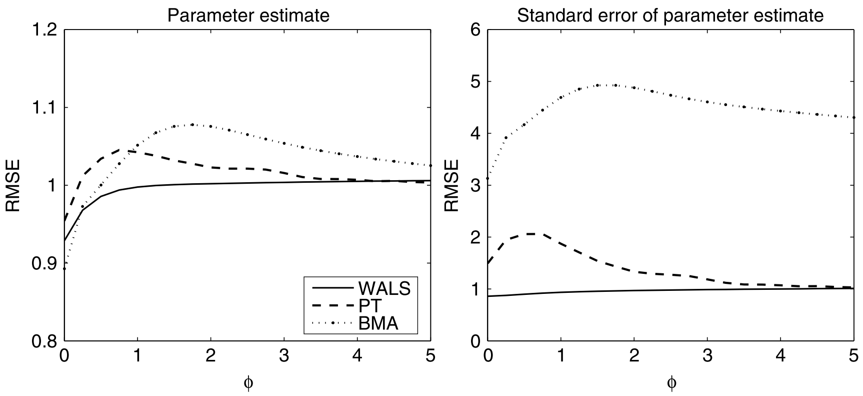

Fig. 4. RMSE comparisons for GARV.

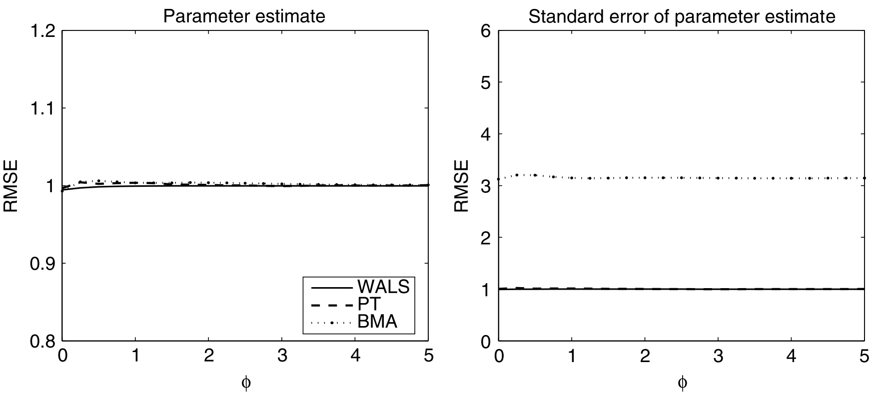

Fig. 5. RMSE comparisons for INDV.

We next turn to the comparisons of the standard errors of the estimates, again in terms of the RMSE (right panels). Here, the results are strongly in favor of WALS. In nearly every case the WALS standard errors of the estimates have smaller RMSE, often by a large margin, than the RMSEs of the BMA and PT estimates. The only exception occurs with the parameters of *LFLOOR* and *INDV* where the pretest standard error of the estimates achieves roughly the same RMSE as the WALS standard error. It is rather striking that the BMA estimator of the standard error performs very poorly in all cases, with a RMSE of three or four times that of WALS. The PT estimator generally performs worse than WALS, though occasionally it can result in a RMSE very close to or slightly better than the RMSE of WALS.

To complete our Monte Carlo experiment we consider the out-of-sample forecasting power of WALS compared with the three competing estimators. We estimate the model described in the previous section by WALS, BMA, pretest, and ML based on the $n_1 = 375$ pre-2007 transactions, Using the estimated models we then construct predictions of y for the remaining $n_2 = 185$ observations in 2007. Let $\hat{y}_i$ be the predicted value of y_i, the ith record. Our evaluations of forecast accuracy are

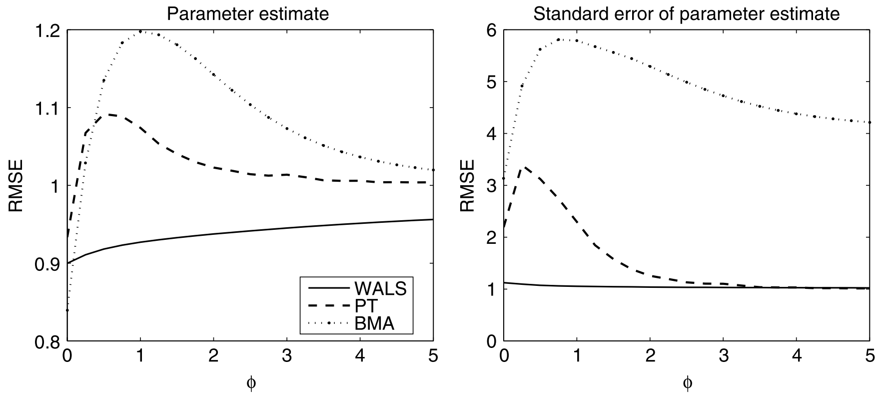

Fig. 6. RMSE comparisons for SEAVF.

Table 3
Predictive accuracy.

	RMSE($\hat{y}$)	BIAS($\hat{y}$)	SE($\hat{y}$)
ML	0.1302	−0.0788	0.1036
WALS	0.1279	−0.0781	0.1012
PT	0.1295	−0.0776	0.1036
BMA	0.1295	−0.0786	0.1030

based on

$$\text{BIAS}(\hat{y}) = \frac{1}{n_2}\sum_{i=1}^{n_2}(\hat{y}_i - y_i), \qquad \text{RMSE}(\hat{y}) = \sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}(\hat{y}_i - y_i)^2},$$

and

$$\text{SE}(\hat{y}) = \sqrt{\text{RMSE}^2(\hat{y}) - \text{BIAS}^2(\hat{y})}.$$

Again, the PT estimator is based on the *stepwisefit* routine in Matlab; the procedure selects *SEAVM*, *SEAVS*, *MONV* and *STRN* in addition to the focus regressors. The results are summarized in Table 3. We see that WALS is the clear favorite while the ML estimator is the least preferred.

5. Conclusions

We have shown in this paper that model averaging can be used as a general method of estimation, not only when the number of regressors is large (as is typically the case in growth studies) but also when the number of regressors is small (as in our application). Since we do not ask which model is the 'true' model, no answer to this question is provided. Instead we ask which statistic is 'best' to estimate the parameters of interest. Model averaging does provide an answer to this question and this answer has attractive features. One feature is that we can leave behind the uncomfortable discontinuity caused by the t-ratio when used in the traditional manner. Instead of making one decision when $t = 1.95$ and a different decision when $t = 1.97$, the t-ratio (and other diagnostics) now provides *continuous* information: if $t = 1.97$ we are a little more confident about this piece of information than when $t = 1.95$. Another feature is that the problem of pretesting is resolved. Model selection and estimation are now treated as one joint activity and the properties of the estimators reflect this.

More specifically, this paper has attempted to do three things. First, we extend the equivalence theorem and the mechanism of WALS estimation to nonspherical disturbances, and obtain the bias and variance of the WALS estimators for this case. Second, we apply the theory to the Hong Kong housing market, and estimate the determinants of apartment prices in the 'South Horizon' housing estate during 2004–2007. Third, and most important, we perform simulations to see how our proposed WALS estimator compares with the Bayesian model average and pretest estimators. There are, of course, other candidates for model averaging, in particular ridge regression, which shrinks the coefficient vector towards zero and has a Bayesian interpretation, and the recently proposed estimator by Hansen (2007), based on the Mallows criterion. It remains for future research to include these alternative estimators in a more comprehensive comparison.

The simulations show that the WALS method yields generally more accurate estimates than do the BMA and pretest estimators, and that the improvement in accuracy can be very substantial. Although our example only deals with models with heteroskedastic disturbances, the theory can be applied straightforwardly to linear models with autocorrelated, GARCH, and any other form of nonspherical disturbances. WALS estimation should thus be considered as a serious alternative to traditional (pretest) estimation and BMA. An added but nontrivial bonus is that computing time for WALS is negligible, even when the number of regressors is large.

Acknowledgements

We are grateful to the editor in charge and two referees for their positive and constructive comments. Wan's work was partially supported by a General Research Fund from the Hong Kong Research Grants Council (No. City-U 102709).

References

Adkins, L.C., 1996. Prior information in regression: to choose or not to choose. Journal of Statistical Computation and Simulation 55, 31–48.
Adkins, L.C., Eells, J.B., 1995. Improved estimators of energy models. Energy Economics 17, 15–25.
Bao, H.X.H., Wan, A.T.K., 2004. On the use of spline smoothing in estimating hedonic housing price models: empirical evidence using Hong Kong data. Real Estate Economics 32, 487–507.
Bao, H.X.H., Wan, A.T.K., 2007. Improved estimators of hedonic housing price models. Journal of Real Estate Research 29, 267–301.
Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 47, 1287–1294.
Buchholz, A., Holländer, N., Sauerbrei, W., 2008. On properties of predictors derived with a two-step bootstrap model averaging approach—a simulation study in the linear regression model. Computational Statistics & Data Analysis 52, 2778–2793.
Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.
Claeskens, G., Croux, C., Van Kerckhoven, J., 2006. Variable selection for logistic regression using a prediction focussed information criterion. Biometrics 62, 972–979.
Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge, UK.
Danilov, D., Magnus, J.R., 2004a. On the harm that ignoring pretesting can cause. Journal of Econometrics 122, 27–46.
Danilov, D., Magnus, J.R., 2004b. Forecast accuracy after pretesting with an application to the stock market. Journal of Forecasting 23, 251–274.
Durlauf, S.N., Johnson, P.A., Temple, J.R.W., 2005. Growth econometrics. In: Aghion, P., Durlauf, S.N. (Eds.), Handbook of Economic Growth. North Holland, Amsterdam, pp. 555–677.
Global Property Guide, 2008. Most expensive cities in 2008. http://www.globalpropertyguide.com/investment-analysis/Most-expensive-cities-in-2008.
George, E.I., 1986. Minimax multiple shrinkage estimation. Annals of Statistics 14, 188–205.
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.
Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial (with discussion). Statistical Science 14, 382–417.
Hui, E.C.M., Yue, S., 2006. Housing price bubbles in Hong Kong, Beijing and Shanghai: a comparative study. Journal of Real Estate Finance and Economics 33, 299–327.
Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York.
Leeb, H., 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., Pötscher, B.M., 2005. Model selection and inference: facts and fiction. Econometric Theory 21, 21–59.
Leeb, H., Pötscher, B.M., 2006. Can one estimate the conditional distribution of post-model-selection estimators? The Annals of Statistics 34, 2554–2591.
Leeb, H., Pötscher, B.M., 2008. Can one estimate the unconditional distribution of post-model-selection estimators? Econometric Theory 24, 338–376.
Magnus, J.R., 2002. Estimation of the mean of a univariate normal distribution with known variance. Econometrics Journal 5, 225–236.
Magnus, J.R., Durbin, J., 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. Econometrica 67, 639–643.
Magnus, J.R., Powell, O., Prüfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. Journal of Econometrics 154, 139–153.
Ouysse, R., Kohn, R., 2010. Bayesian variable selection and model averaging in the arbitrage pricing theory model. Computational Statistics & Data Analysis 54, 3249–3268.
Peña, D., Redondas, D., 2006. Bayesian curve estimation by model averaging. Computational Statistics & Data Analysis 50, 688–709.
Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 179–191.
Schomaker, M., Wan, A.T.K., Heumann, C., 2010. Frequentist model averaging with missing observations. Computational Statistics & Data Analysis 54, 3336–3347.
Wan, A.T.K., 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., Zou, G., 2010. Least squares model averaging by Mallows criterion. Journal of Econometrics 156, 277–283.
Wong, J.T.Y., Hui, E.C.M., Seabrooke, W., Raftery, J., 2005. A study of the Hong Kong property market: housing price expectations. Construction Management and Economics 23, 757–765.
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.
Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? Journal of the American Statistical Association 100, 1202–1214.