University of Victoria | Economics

Econometrics Working Paper EWP1701

ISSN 1485-6441

MODEL AVERAGING OLS AND 2SLS: AN APPLICATION OF THE WALS PROCEDURE

Judith Anne Clarke
Department of Economics, University of Victoria
Victoria, B.C., Canada V8W 2Y2

June 2017

Summary

More commonly, applied and theoretical researchers are examining model averaging as a tool when considering estimation of regression models. Weighted-average least squares (WALS), originally proposed by Magnus and Durbin (1999, *Econometrica)* within the framework of estimating some of the parameters of a linear regression model when other coefficients are of no interest, is one such model averaging method with their proposed approach being a Bayesian combination of frequentist ordinary least squares and restricted least squares estimators. We generalize their work, along with that of other researchers, to consider averaging ordinary least squares (OLS) and two stage least squares (2SLS) estimators when possibly one or more regressors are endogenous. We derive asymptotic properties of our weighted OLS and 2SLS estimator under a local misspecification framework, showing that results from the existing WALS literature apply equally well to our case. In particular, determining the optimal weight function reduces to the problem of estimating the mean of a normally distributed random variate, which is unrelated to the details specific to the regression model of interest, including the extent of correlation between the explanatory variable(s) and the error term. We illustrate our findings with two examples. The first example, from a commonly adopted econometrics textbook, considers returns to schooling, and the second case is a growth regression application, which examines whether religion assists in explaining disparities in cross-country economic growth.

Keywords: Model averaging; least squares; two stage least squares; priors; instrumental variables.

JEL Classifications: *C11, C13, C26, C51, C52*

Author Contact:
Judith Anne Clarke, Department of Economics, University of Victoria, P.O. Box 1700, STN CSC, Victoria, BC, Canada V8W 2Y2; E-mail: jaclarke@uvic.ca; Voice: (250) 721-8542; FAX: (250) 721-6214

1. Introduction and Preliminaries

We consider a structural equation

$$y = X_1\beta_1 + X_2\beta_2 + u, \;\; X_2 = \mu + v \tag{1}$$

that is part of a system of simultaneous equations. The setup of this equation is as follows: the vectors y and u are $n-$dimensional, v is a $(n \times K_2)$ random matrix; the full-rank regressor matrices X_1 and X_2 are, respectively, $(n \times K_1)$ and $(n \times K_2)$ with $X = [X_1\; X_2]$ and $K = K_1 + K_2$; the $(K_1 \times 1)$ and $(K_2 \times 1)$ vectors of unknown parameters are, respectively, β_1 and β_2; and μ is a $(n \times K_2)$ nonstochastic matrix. We can equivalently write model (1) as $y = X\beta + u$, where $\beta' = [\beta_1' \; \beta_2']$. Our assumptions and notation are:

Assumption 1: $[u_i \quad v_i] \sim i.i.d.\ N\left(0_{1\times(K_2+1)}, \Omega\right), i = 1, \ldots, n,$ where $\Omega = \begin{bmatrix} \sigma^2 & \delta' \\ \delta & \Omega_0 \end{bmatrix}$.

Assumption 2: The regressors in the matrix X_1 are exogenous, independent variables with $E(u|X_1) = 0_{n\times 1}$ & $E(X_1'u) = 0_{K_1\times 1}$.

Assumption 3: The $(K \times K)$ matrix $Q_{XX} = plim(X'X/n)$ exists. Appropriate submatrices are denoted by $Q_{X_1X_1}$, $Q_{X_1X_2}$, $Q_{X_2X_2}$ and $Q_{X_2X_1}$.

Assumption 4: The system provides a $(n \times L_2)$ matrix of instruments Z_2 such that $L_2 \geq K_2$ and

(i) $plim(Z_2'X_2/n) = Q_{Z_2X_2}$ exists with rank K_2. We denote $Q_{Z_2X_2}'$ as $Q_{X_2Z_2}$.
(ii) $plim(Z_2'u/n) = 0_{L_2\times 1}$, $plim(Z_2'v/n) = 0_{L_2\times L_2}$.

Assumption 5: Let $L = K_1 + L_2$ and $Z = [X_1\; Z_2]$ be the instrument matrix. We assume that:

(i) $plim(Z'X/n) = Q_{ZX}$ exists with rank K,
(ii) $plim(Z'Z/n) = Q_{ZZ}$ exists with rank L.

Some relevant submatrices, to be used in a later section, are given by Q_{X_2Z}, Q_{X_1Z}, Q_{ZX_2} and Q_{ZX_1}.

These assumptions imply some possible correlation between the errors and the explanatory variables in X_2 with there being no correlation when $\delta = 0$. Explicitly, we have

$$plim(X'u/n) = \begin{bmatrix} plim(X_1'u/n) \\ plim(X_2'u/n) \end{bmatrix} = \begin{bmatrix} 0_{K_1\times 1} \\ \delta \end{bmatrix} = \xi.$$

The instrument matrix Z is employed to form the two-stage least squares (2SLS) estimator. Then, the ordinary least squares (OLS) and 2SLS estimators of β are, respectively:

$$\hat{\beta}_{(i)} = (X'P_iX)^{-1}X'P_iy \quad ; i = 1,2 \tag{2}$$

where $P_1 = I_n$ and $P_2 = Z(Z'Z)^{-1}Z'$; we denote $\hat{\beta}_{(i,j)}$ as the estimator of β_j ($j = 1,2$) using OLS ($i = 1$) and 2SLS ($i = 2$). With $\delta \neq 0$, the OLS estimator is both biased and inconsistent[1], whereas the 2SLS estimator is (usually) biased but consistent. Specifically, it is straightforward to show that the limiting distributions of $\hat{\beta}_{(1)}$ and $\hat{\beta}_{(2)}$ are:

$$\hat{\beta}_{(1)} \overset{a}{\sim} N\left(\beta + Q_{XX}^{-1}\xi, \frac{\sigma^2}{n} Q_{XX}^{-1}\right) \tag{3a}$$

and

$$\hat{\beta}_{(2)} \overset{a}{\sim} N\left(\beta, \frac{\sigma^2}{n} B^{-1}\right), \tag{3b}$$

where $B = Q_{ZX}' Q_{ZZ}^{-1} Q_{ZX}$. The asymptotic covariance matrix of $\hat{\beta}_{(1)}$ is never "larger", in the matrix sense, than that of $\hat{\beta}_{(2)}$ and will usually be "smaller", unless a regression of X on Z perfectly predicts X. These features of the limiting distributions imply a potential asymptotic mean squared error (or asymptotic risk under quadratic loss) gain of $\hat{\beta}_{(1)}$ over $\hat{\beta}_{(2)}$, despite the inconsistency, dependent on the degree of correlation between X_2 and u, as given by δ.

Traditionally, the goal has been to choose either OLS or 2SLS. From an empirical perspective, this uncertainty about which estimator (model) to adopt is potentially troubling as the estimates we obtain, and so any subsequent conclusions using such estimates, likely depends on which estimator (model) is employed. This leads to a popular strategy[2] of undertaking a preliminary test of whether there is no covariance between X and u:

$$H_0: E(X_2'u) = E(v'u) = n\delta = 0 \quad against\ H_1: n\delta \neq 0. \tag{4}$$

The OLS estimator is unbiased, efficient and consistent under the null hypothesis whereas we are only confident that the 2SLS is consistent when $\delta = 0$. When the null hypothesis is not valid, the 2SLS estimator is still consistent while the OLS estimator is biased and inconsistent but may still have "smaller" variance. Prior testing aims to ensure that a researcher uses the more efficient OLS estimator when endogeneity is not an issue but employs the less efficient 2SLS estimator in the face of large possible risk for OLS when endogeneity covariance is serious. A common route is to apply the so-called Durbin-Wu-Hausman (DWH) (Durbin, 1954; Wu, 1973; Hausman, 1978) test based on the difference ($\hat{\beta}_{(2)}$-$\hat{\beta}_{(1)}$). When the null is valid, the two estimators should only differ due to sampling error, as both estimators are consistent, whereas only the 2SLS estimator is consistent under the alternative hypothesis. One version of the DWH test statistic, the Hausman form, is:

$$DWH = \left(\hat{\beta}_{(2)} - \hat{\beta}_{(1)}\right)' \left[AV\left(\hat{\beta}_{(2)}\right) - AV\left(\hat{\beta}_{(1)}\right)\right]^{-} \left(\hat{\beta}_{(2)} - \hat{\beta}_{(1)}\right) \tag{5}$$

[1] For estimating both β_1 and β_2.

[2] For instance, Baum et al. (2003, p2) state "…the use of IV estimation to address this problem (nonorthogonality) must be balanced against the inevitable loss of efficiency vis-à-vis OLS. It is therefore very useful to have a test of whether or not OLS is inconsistent …".

where $AV(\hat{\beta}_{(i)})$ is the usual asymptotic variance of $\hat{\beta}_{(i)}$ ($i = 1,2$), as provided in expressions (3a) and (3b), and the operator "$^-$" denotes a generalized inverse. Under our assumptions, it can be shown that the limiting null distribution of DWH is central $\chi^2(K_2)$, the distribution usually adopted in practice to generate p-values, as opposed to the exact finite-sample null distribution[3]. Critical values and p-values are often obtained also from an appropriate central F-distribution. Let c be the critical value associated with some chosen nominal significance level α. Adopting $\hat{\beta}_{(1)}$ or $\hat{\beta}_{(2)}$ dependent on the outcome of the test leads to the pretest estimator:

$$\hat{\beta}_{PT} = \hat{\beta}_{(1)} I_{[DWH<c]} + \hat{\beta}_{(2)} I_{[DWH \geq c]} \tag{6}$$

where $I_{[DWH<c]} = 1$ when the sample value of the statistic is less than the critical value, zero otherwise. A similar definition holds for $I_{[DWH \geq c]}$; note that $I_{[DWH \geq c]} = 1 - I_{[DWH<c]}$. Morey (1984) obtains the asymptotic bias and asymptotic risk (under quadratic loss) functions of $\hat{\beta}_{PT}$ showing the common outcome that the pretest estimator is never preferred (in terms of risk) to either of its component estimators[4]. Further, the pretest estimator is asymptotically unbiased (Skeels and Taylor, 1995).

Although the use of $\hat{\beta}_{PT}$ is common, it is well-known that such pretest estimators, being discontinuous and therefore not differentiable, are inadmissible (see, e.g., Judge and Bock, 1978; Magnus, 1999; Magnus, 2002)[5]. An additional, practical concern is that the pretest estimator implies a switch from one estimator (e.g., OLS) to the other estimator (e.g., 2SLS) around the adopted critical value c, a point at which the two component estimators are likely quite similar. Recognizing that the purpose is to find a preferred estimator of β, allowing for uncertainty about some feature of the underlying data generating process, rather than a goal of obtaining the outcome of a preliminary hypothesis test or choosing between one or more models or estimators, has resulted in a wide literature that considers combining component estimators in a smooth, averaging way. The notion is that as each model or estimator provides information about the parameters of interest, it is most likely preferable to use some weighted-average of the component estimators, rather than to employ only one of the estimators in the switching manner implied by pretesting; see, for example, the eloquent discussion in Magnus and De Luca (2016). Averaging in this way incorporates the uncertainty the researcher has regarding the model specification. Non-Bayesian and Bayesian approaches have been proposed, with random and non-random weight functions[6]. A model averaging estimator of β, $\hat{\beta}_{MA}$, for our problem is:

$$\hat{\beta}_{MA} = \lambda \hat{\beta}_{(2)} + (1 - \lambda) \hat{\beta}_{(1)} \tag{7}$$

---

[3] The test is really one of the implications of using two different estimators as opposed to a strict test of endogeneity.
[4] Skeels and Taylor (1995) investigate the finite-sample properties of $\hat{\beta}_{PT}$ providing approximate bias and risk functions using a Taylor series expansion to approximate the bias function and nonparametric methods to evaluate pretest risk, following the approaches of Richardson and Wu (1971) and Gourieroux and Trognon (1984).
[5] We refer the reader to, for instance, Giles and Giles (1993), Danilov and Magnus (2004) and Magnus and De Luca (2016) for discussions on various implications of employing pretest estimators.
[6] There is an extensive literature related to Bayesian model averaging (BMA); e.g., Raftery et al. (1997), Hoeting et al. (1999) and the recent survey of Moral-Benito (2015). See the excellent monograph of Claeskens and Hjort (2008) for a discussion on non-Bayesian/frequentist model averaging approaches, along with, amongst many others, Hjort and Claeskens (2003a), Hansen (2007), Liang et al. (2011) and Moral-Benito (2015).

where the non-negative λ is some continuous, weighting function, possibly specified according to a statistical criteria, along with perhaps prior information. Expression (7) shows that we can view $\hat{\beta}_{MA}$ as a direct, continuous form of the discontinuous, inadmissible pretest estimator, $\hat{\beta}_{PT}$, and highlights that model averaging proceeds in two steps. We first estimate the parameters conditional on the model assumptions (the regressors of concern are endogenous or the relevant explanatory variables are indeed, at least, weakly exogenous). The second stage is to form the model averaging estimator as a weighted-average of these conditional estimators. Of interest is the sampling properties of $\hat{\beta}_{MA}$, along with optimal (according to some specified criterion) choice of λ. Our consideration of the estimator $\hat{\beta}_{MA}$ follows the suggestions of, for example, Morey (1984), Moral-Benito (2014) and Magnus and De Luca (2016). Specifically, Morey (1984, p70) says, following his observations on the risk properties of $\hat{\beta}_{PT}$, "Possibly a more rational and powerful strategy … (is to) employ some weighted linear combination of the OLS and IV estimators …", Moral-Benito (2015, p69) ends his survey paper stating "How to tackle the issue of endogenous regressors in the model averaging framework is an interesting line of open research. … Allowing for endogenous regressors in the FMA[7] approach could be an interesting topic for future research," and Magnus and De Luca (2016, p138) remark that an important addition to their work would be "a WALS version of instrumental variables or two-stage least squares". Our work begins this research.

We average the OLS and 2SLS estimators using weighted-average least squares (WALS), a method originally proposed by Magnus and Durbin (1999) in the context of deciding on a preferred set of explanatory variables in a classical linear regression model, and further examined by Danilov and Magnus (2004), Danilov (2005), Zou et al. (2007), Clarke (2008), Magnus et al. (2010), De Luca and Magnus (2011) to name but a few. We write our weighted OLS and 2SLS estimator, based on WALS, as $\hat{\beta}_W$, to distinguish this estimator from some other, feasible model averaging estimator, $\hat{\beta}_{MA}$. We show that results from the already existing WALS literature, despite the seemingly quite different setting of that body of work, can be readily extended to obtain the asymptotic bias, variance and risk (under quadratic loss) functions of $\hat{\beta}_W$, under a structure of local misspecification. This outcome implies that the optimal (in terms of asymptotic risk under quadratic loss) $\hat{\beta}_W$ is determined solely by establishing the optimal estimator of the mean of a normal random variate, unrelated to the specifics of the regression model (1), including the degree of correlation between X and u. Our findings also mean that prior research on selecting λ, dependent on adopted criterion, apply to optimally combining the OLS and 2SLS estimators, at least asymptotically, to obtain preferred coefficient estimates.

Although we are novel in applying WALS in the context of weighting the frequentist OLS and 2SLS estimators, we note that other researchers have considered Bayesian model averaging of 2SLS/LIML (limited information maximum likelihood) estimators, including Durlauf et al. (2008, 2012) and Lenkoski et al. (2014). Concern about the selection of endogenous and exogenous regressors is the focus of the work of Durlauf et al. (2008) when exploring empirical evidence for a number of growth theories. They suggest that their approach can be regarded as a

[7] Frequentist Model Averaging

frequentist/Bayes hybrid method that averages frequentist 2SLS estimates from each of the considered models with weights regarded as posterior probabilities that link prior weights with complexity-penalized goodness of fit measures. Lack of formal justification for their choice of weights is noted by the authors. A similar hybrid Bayesian model averaging approach is adopted by Durlauf et al. (2012) in their re-examination of the robustness of Barro and McCleary's (2003) finding that some facets of religious dogmas are relevant for economic growth. Taking account of model uncertainty, their results highlight the lack of robustness of findings based on models that assume to be the true specification. Extending such BMA methods of mixing 2SLS estimators when there is additionally uncertainty with regard to the choice of instruments is the topic of interest for Lenkoski et al. (2014).

Our work of examining mixing OLS and 2SLS estimators within a WALS framework is organized as follows. In Section 2 we outline key papers exploring WALS, as applied to model averaging in the context of the presence of uncertain auxiliary variables, the question of which has dominated this body of work. We also provide salient details on considered extensions and some Monte Carlo (MC) experiments. We derive further generalizations to the existing WALS literature in Section 3, which we then apply, in Section 4, to examine the WALS estimator that averages the OLS and 2SLS estimators. Specifically, we show that the asymptotic bias and asymptotic risk (under quadratic loss) functions of the weighted estimator $\hat{\beta}_W$ are readily obtained from the extensions of the theory presented in Section 3. In Section 5, we draw on existing empirical studies to demonstrate our results, illustrating with two applications, and Section 6 concludes.

2. WALS: Some Research

Predominantly, the current WALS literature examines a classical linear regression model for which uncertainty exists on whether to include one or more of a set of possible auxiliary variables, with the other explanatory variables (so-called focus variables) being deemed required in the specification. Specifically, consider the linear model:

$$y = W_1\gamma_1 + W_2\gamma_2 + \varepsilon = W\gamma + \varepsilon,\ \ \varepsilon \sim N(0, \sigma_\varepsilon^2 I_{n\times n}) \quad (8)$$

where the vectors y and ε are $n-$dimensional, the full-rank regressor matrices W_1 and W_2 are, respectively, $(n\times k_1)$ and $(n\times k_2)$ with $W = [W_1\ W_2]$ and $k = k_1 + k_2$ and the $(k_1 \times 1)$ and $(k_2 \times 1)$ vectors of unknown parameters are, respectively, γ_1 and γ_2; let $\gamma' = [\gamma_1'\ \gamma_2']$. The columns in W_1 contain the "focus" variables, those regressors that must be in the specification, whereas inclusion of the "auxiliary" variables in W_2 is less certain. Key interest is in the estimation of γ_1, the "focus" parameters; one or more of the regressors in W_2 are included only to lead to a "better" estimator of γ_1. This framework implies a model space, say Υ, which comprises the set of submodels of (8) that include all focus variables in W_1 and some (one or more) of the auxiliary variables in W_2; there are 2^{k_2} submodels in this space with the i^{th} submodel Υ_i, being model (8) subject to the restriction $H_i'\gamma_2 = 0$, where H_i is a selection matrix of dimension $(k_2 \times f_i)$ that picks out the variables in W_2 excluded to form submodel Υ_i, $i = 1, \ldots, 2^{k_2}$; $0 \le f_i \le k_2$. The goal is to examine the sampling properties of the model averaging, WALS, estimators of γ_1 and γ_2 defined respectively as:

$$g_1 = \sum_{i=1}^{2^{k_2}} \eta_{(i)} \hat{\gamma}_{1(i)} \quad \& \quad g_2 = \sum_{i=1}^{2^{k_2}} \eta_{(i)} \hat{\gamma}_{2(i)} \tag{9}$$

where $\hat{\gamma}_{1(i)}$ and $\hat{\gamma}_{2(i)}$ are the OLS estimators of γ_1 and γ_2 for submodel $\Upsilon_i, i = 1, \dots, 2^{k_2}$. The model weights are given by $\eta_{(i)}$, assumed to satisfy $0 \le \eta_{(i)} \le 1$; $\sum_i \eta_{(i)} = 1$; and $\eta_{(i)} = \eta_{(i)}(C_{W_1} y)$, where $C_{W_1} = I_{n\times n} - W_1(W_1'W_1)^{-1}W_1'$; see, e.g., Magnus and De Luca (2016, p.123) for a discussion on the latter condition[8]. Let $\theta = (W_2'C_{W_1}W_2)^{\frac{1}{2}}\gamma_2$ and $\hat{\theta} = (W_2'C_{W_1}W_2)^{-\frac{1}{2}}W_2'C_{W_1}y$; note that $\hat{\theta}\sim N(\theta, \sigma_\varepsilon^2 I_{k_2\times k_2})$. A key result shown from this body of research (specifically Magnus and Durbin, 1999; Danilov and Magnus, 2004; Magnus et al., 2010, Magnus and DeLuca, 2016) is that $MSE(g_1)$ and $MSE(g_2)$ are determined solely by $MSE(\tilde{\theta})$, where $\tilde{\theta} = T\hat{\theta}$ with $T = \sum_{i=1}^{2^{k_2}} \eta_{(i)} T_i$, $T_i = I_{k_2\times k_2} - t_i$, $t_i = (W_2'C_{W_1}W_2)^{-\frac{1}{2}} H_i \times \left[H_i'(W_2'C_{W_2}W_2)^{-1}H_i\right]^{-1} H_i'(W_2'C_{W_1}W_2)^{-\frac{1}{2}}$, where mean squared error ($MSE(.)$) is risk under squared error loss. This fundamental result implies that the mean squared error of the complicated estimators g_1 and g_2 is determined by the mean squared error of the simpler estimator $\tilde{\theta} = T\hat{\theta}$ of θ. In particular, the question of the "best" weight function for the WALS estimators g_1 and g_2 is decided on by ascertaining the "best" weight function for the weighted estimator $\tilde{\theta}$ of θ, the mean vector of a normal distribution – a problem that does not depend on the specifics of the underlying regression model (8).

Various studies have extended the seminal works of Magnus and Durbin (1999) and Danilov and Magnus (2004). We detail a few of them here. Zou et al. (2007), within the framework of Magnus and Durbin (1999), generalize results to the large-sample non-normal errors case, presenting asymptotic bias and risk functions. They show that the findings of Magnus and Durbin still hold; i.e., the question of the optimal choice of weight function for the WALS estimator is addressed by determining how best to estimate the mean from a single normal observation. As well as broadening the applicability of Magnus and Durbin's conclusions to the large-sample non-normal case, Zou et al. examine the issue of the optimal weight function, including the "neutral" Laplace Bayesian approach of Magnus (2002) and the James and Stein (1960) type weight studied by Kim and White (2001). In addition, these authors explore non-random weight functions, along the lines considered by, amongst others, Kim and White (2001) and Hjort and Claeskens (2003a).

Clarke (2008) considers estimation of the full parameter vector γ in model (8), when there is uncertainty about a set of J exact linear restrictions about γ, say $R\gamma = r$, where R $(J \times k)$ and r $(J \times 1)$ are nonstochastic with the matrix R being of full row-rank, $J < k$. Her framework covers, but is not limited to, the case of estimating some coefficients when others are not of interest that had dominated research. She shows that the theorems of Magnus and Durbin (1999)

[8] In addition, Magnus and De Luca (2016, pp. 124-125) present discussion on the question of limiting weights to lie between zero and one.

and Danilov and Magnus (2004) hold equally well to this commonly applied broader setup; the optimal MSE weighted estimator of γ is determined solely by the weight function for optimally estimating the mean vector of an uncorrelated, homoskedastic normally distributed variate. Her work and the asymptotic approach taken by Zou et al. (2007) are useful for our research.

Irrespective of the elegance of the findings from these aforementioned papers, at first glance these results suggest that determining the optimal WALS estimators, g_1 and g_2, implies ascertaining 2^{k_2} different $\eta_{(i)}'s$. This is not the case – of importance is resolving the question of the matrix T, a symmetric $(k_2 \times k_2)$ matrix with only $k_2(k_2+1)/2$ distinct elements. Indeed, if certain orthogonalizations are applied to the model prior to estimation, this matrix can be rendered diagonal so that only k_2 elements need be established – a significantly smaller task than might originally be suspected, and rendering WALS readily useable in applications, especially when compared with commonly adopted BMA approaches. Considering a semi-orthogonal transformation of the auxiliary regressors, within the setup of Danilov and Magnus's (2004) "focus" and "auxiliary" variables modelling exercise, is one goal of Magnus et al. (2010), along with the question of how then to practically proceed; see, also, for instance, Magnus (2002), Zou et al. (2007), Magnus et al. (2010), Kumar and Magnus (2013), Magnus and De Luca (2016).

Several studies expand on the initial WALS framework. Allowing for nonspherical disturbances partly motivates Magnus et al. (2011) in their extension of the theorems of Magnus and Durbin (1999) and Danilov and Magnus (2004), again within the setup of "focus" and "auxiliary" variables. After semi-orthogonalizing the model and considering appropriate unrestricted and restricted generalized least squares estimators, they show that the theorems of Magnus and Durbin (1999) and Danilov and Magnus (2004) (as well as Magnus et al., 2010) can be extended to the nonspherical disturbances case. An interesting tiered approach is proposed by Magnus and Wang (2014), so-called hierarchical WALS, whereby model averaging is performed over uncertain explanatory variable concepts and the different ways one might measure the variable concept. De Lucca et al. (2017) broaden the WALS approach, still within the framework of "focus" and "auxiliary" regressors, to consider the specification of the linear predictor in the class of generalized linear models.

Some researchers have undertaken Monte Carlo (MC) experiments to ascertain the properties of WALS estimators. For instance, Magnus et al. (2011) use their real-world data to setup a MC experiment that compares the root mean squared error of their considered WALS, BMA and pretest estimators, with their simulations showing that their WALS estimator is typically more accurate. Poghosyan and Magnus (2012) undertake a small MC experiment based on their facto-based dynamic models using Armenian data, examining the performance of their WALS and BMA estimators. They find that, although the WALS method usually outperforms, in terms of root mean squared error, the BMA estimator, the differences can be quite small. Using the first two waves of the Survey of Health, Ageing and Retirement in Europe as a base for the design of their MC study, De Lucca et al. (2017) investigate the finite sample properties of their WALS estimator.

In addition to exploring properties of WALS estimators, several studies have employed WALS estimators in applied research. Magnus et al. (2010), who contrast WALS and BMA estimates, along with a general-to-specific pretesting strategy and those from fully unrestricted and restricted versions, consider modelling growth in per capita GDP between 1960 and 1996 for a cross-section of 74 countries[9]. They split their 13 possible regressors into two groups: the focus variables and auxiliary variables, examining several frameworks; e.g., one specification (a neo-classical growth model) contains six focus and four auxiliary regressors, whereas another (an endogenous growth model) consists of nine focus and four auxiliary variables. On the basis of this work, the authors suggest that their WALS findings clarify some issues on growth determinants and theories; in particular, that robust growth models should include information on institutions, along with neoclassical growth variables.

An hedonic housing price model for apartments using data from the Hong Kong real estate market is estimated by Magnus et al. (2011), contrasting outcomes of the WALS estimator[10], a BMA estimator[11] and a stepwise fit pretest estimator[12]. WALS[13] and BMA estimators, similar to those considered by Magnus et al. (2011), are compared for four factor-based dynamic models using quarterly Armenian data in Poghosyan and Magnus (2012). The concept of hierarchical WALS is applied by Magnus and Wang (2014) to re-examine data from Sala-i-Martin et al. (2004), who consider the effects of various growth factors. One key aim of the study of Magnus and Wang (2014) is to illustrate how WALS is helpful in handling the common situation of more than one variable being available for a factor; a two-level strategy is proposed whereby the first layer considers which concepts are best included in the model with the second layer exploring the question of how best to represent the concept with the various available variables. Adopting the form of WALS examined by Magnus et al. (2010) and De Luca and Magnus (2011), with a Laplace prior to form the weight function, Magnus and Wang's results suggest that the signs of the hierarchical WALS estimates are more intuitive and robust than those from traditional methods. These empirical works highlight that WALS is a viable estimation approach for practitioners.

3. Main General Results

Our goal in this section is to generalize the results of Magnus and Durbin (1999), Danilov and Magnus (2004), Zou et al. (2007) and Clarke (2008). Specifically, consider estimation of model (8) in the face of uncertainty about $J\ (< k)$ exact linear restrictions on γ, written as $R\gamma = r$, where $R\ (J \times k)$ and $r\ (J \times 1)$ are nonstochastic with R of full-row rank. With $D = (W'W)$, the unrestricted least squares estimator of γ is denoted by $g_u = D^{-1}W'y$, and we write the restricted least squares (RLS) estimator by $g_r = g_u - D^{-1}R'[RD^{-1}R']^{-1}(Rg_u - r)$. We are interested in the

---

[9] Their primary source of data is from Sala-i-Martin et al. (2004). In addition to their exploration of 74 countries, the authors analyze Sala-i-Martin et al.'s (2004) full data set.
[10] Employing a weight function based on the method suggested by Magnus et al. (2010).
[11] With priors based on the normal distribution.
[12] The authors adopt a model with six "focus" variables and six "auxiliary" variables. The stepwise fit pretest strategy starts with the model with the six focus variables, adding an auxiliary regressor based on statistical significance.
[13] The version of WALS applied is that detailed in Magnus et al. (2010), using the weighting scheme outlined by De Luca and Magnus (2011) to ensure scale-independent estimates.

asymptotic bias and risk (under quadratic loss) functions of the weighted combination of g_u and g_r:

$$g = \eta(\hat{\phi})g_u + \left(1 - \eta(\hat{\phi})\right) g_r \tag{10}$$

where $\eta(.)$ is a continuous, random function, dependent on $\hat{\phi} = [RD^{-1}R']^{-\frac{1}{2}}(Rg_u - r)$, an estimator of ϕ, defined below, and perhaps some other arguments (e.g., an estimator of σ_ε^2). We assume that σ_ε^2 is known but, given our asymptotic focus, this is not a limiting assumption should a consistent estimator of σ_ε^2 be available. We later discuss estimation of σ_ε^2.

Following Zou et al. (2007), let γ_0 be any solution to $R\gamma = r$, and let $\gamma = \gamma_0 + \psi/\sqrt{n}$, so that the elements of $\psi, \psi_1, \ldots, \psi_k$, denote departures from the restrictions of interest. This setup implies that we are examining asymptotic properties of the WALS estimator for small departures from the beliefs regarding the parameters, which we hope does not limit the applicability of our results. Then, it is straightforward to show that:

$$\hat{\phi} = \left[RD^{*-1}R'\right]^{-1/2}R\psi + \left[RD^{*-1}R'\right]^{-1/2}RD^{*-1}V_1,$$

$$\sqrt{n}(g_u - \gamma) = D^{*-1}V_1,$$

$$\sqrt{n}(g_r - \gamma) = -D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\psi + \left[I_{k\times k} - D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\right]D^{*-1}V_1,$$

$$\sqrt{n}(g_u - g_r) = D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\left(D^{*-1}V_1 + \psi\right)$$

where $V_1 = \sqrt{n}A$ with $A = W'\varepsilon/n$, and $D^* = D/n$. Let $M_{WW} = plim\ D^*$, which we assume is a nonstochastic positive-definite matrix. In addition, we assume that the regressors are such that $plimA = 0_{k\times 1}$. Now, with $F = M_{WW}^{-1}R'[RM_{WW}^{-1}R']^{-1/2}$, $\Psi = M_{WW}^{-1} - FF'$, writing ϕ as $\phi = [RM_{WW}^{-1}R']^{-1/2}R\psi$ and noting that $V_1 \overset{d}{\rightarrow} U_1 \sim N(0, \sigma_\varepsilon^2 M_{WW})$, we have:

$$\begin{pmatrix} \hat{\phi} \\ \sqrt{n}(g_r - \gamma) \\ \sqrt{n}(g_u - \gamma) \\ \sqrt{n}(g_u - g_r) \end{pmatrix} \overset{d}{\rightarrow} U \sim N(\kappa, \Sigma) \tag{11}$$

where $U = \begin{pmatrix} U_2 \\ U_3 \\ U_4 \\ U_5 \end{pmatrix} = \begin{pmatrix} \phi + F'U_1 \\ -F\phi + \Psi U_1 \\ M_{WW}^{-1}U_1 \\ FU_2 \end{pmatrix}$, $\kappa = \begin{pmatrix} \phi \\ -F\phi \\ 0_{k\times 1} \\ F\phi \end{pmatrix}$ and

$$\Sigma = \sigma_\varepsilon^2 \begin{pmatrix} I_{J\times J} & 0_{J\times k} & F' & F' \\ 0_{k\times J} & \Psi & \Psi & 0_{k\times k} \\ F & \Psi & M_{WW}^{-1} & FF' \\ F & 0_{k\times k} & FF' & FF' \end{pmatrix}.$$

Using these various expressions, we have:

$$
\begin{aligned}
\sqrt{n}(g-\gamma) &= \eta(\hat{\phi})\left(\sqrt{n}(g_u-g_r)\right)+\sqrt{n}(g_r-\gamma)\\
&\overset{d}{\rightarrow} \eta(\phi)U_5+U_3\\
&= \eta(\phi)(FU_2)+U_3. \qquad (12)
\end{aligned}
$$

Denoting $\Xi = \eta(\phi)(FU_2)+U_3$, the asymptotic bias function for g is:

$$
\begin{aligned}
AB(g,\gamma) &= E(\Xi)\\
&= E(\eta(\phi)(FU_2)+U_3)\\
&= FE(\eta(\phi)U_2-\phi).
\end{aligned}
$$

It then follows that the asymptotic variance of g is:

$$
\begin{aligned}
AV(g,\gamma) &= AV(\Xi)\\
&= AV(\eta(\phi)(FU_2)+U_3)\\
&= F[AV(\eta(\phi)U_2)]F'+\sigma_\varepsilon^2\Psi.
\end{aligned}
$$

In terms of asymptotic risk, we follow, for example, Zou et al. (2007), along with many others, in considering the quadratic loss function $L_1(g,\gamma) = \left(\sqrt{n}(g-\gamma)\right)' \Pi_n \left(\sqrt{n}(g-\gamma)\right) = \left\|\sqrt{n}(g-\gamma)\right\|_{\Pi_n}^2$ and the matrix loss function $L_2(g,\gamma) = (n(g-\gamma)(g-\gamma)')$, where Π_n is a random, positive semi-definite $(k \times k)$ matrix. Common choices of Π_n include an identity matrix, which leads to the sum of squared error losses, and $\Pi_n = W'W$ results in the so-called predictive loss function. Using these definitions, with $\frac{1}{n}\Pi_n \overset{p}{\rightarrow} \Pi$, our results imply:

$$
L_1(g,\gamma) \overset{d}{\rightarrow} \|\Xi\|_\Pi^2 = \|\eta(\phi)(FU_2)+U_3\|_\Pi^2
$$

and

$$
L_2(g,\gamma) \overset{d}{\rightarrow} \Xi\Xi' = (\eta(\phi)(FU_2)+U_3)(\eta(\phi)(FU_2)+U_3)',
$$

which leads to the following asymptotic risk functions:

$$
\begin{aligned}
AR_1(g,\gamma) &= E\|\eta(\phi)(FU_2)+U_3\|_\Pi^2\\
&= E\|\eta(\phi)(FU_2)-F\phi\|_\Pi^2+E\|U_3+F\phi\|_\Pi^2\\
&= E\|\eta(\phi)(FU_2)-F\phi\|_\Pi^2+E\|\Psi U_1\|_\Pi^2\\
&= E\|\eta(\phi)U_2-\phi\|_{F'\Pi F}^2+\sigma_\varepsilon^2 tr(\Pi\Psi) \qquad (13)
\end{aligned}
$$

and

$$
AR_2(g,\gamma) = E\left((\eta(\phi)(FU_2)+U_3)(\eta(\phi)(FU_2)+U_3)'\right)
$$

$$= E\big((\eta(\phi)(FU_2) - F\phi)(\eta(\phi)(FU_2) - F\phi)'\big)$$
$$+E(U_3 + F\phi)(U_3 + F\phi)')$$
$$= F\big(E((\eta(\phi)U_2 - \phi)(\eta(\phi)U_2 - \phi)')\big)F' + \sigma_\varepsilon^2\Psi. \quad (14)$$

These outcomes give rise to Theorem 1.

Theorem 1: Under the framework of this section, the asymptotic bias, variance and risk functions of the WALS estimator are:

$$AB(g,\gamma) = F[AB(\eta(\phi)U_2, \phi)],$$
$$AV(g,\gamma) = F[AV(\eta(\phi)U_2)]F' + \sigma_\varepsilon^2\Psi,$$
$$AR_1(g,\gamma) = E\|\eta(\phi)U_2 - \phi\|^2_{\mathrm{F'\Pi F}} + \sigma_\varepsilon^2 tr(\Pi\Psi),$$

and

$$AR_2(g,\gamma) = F[AR_2(\eta(\phi)U_2, \phi)]F' + \sigma_\varepsilon^2\Psi$$

where the $(J \times 1)$ vector U_2 is distributed as $U_2 \sim N(\phi, \sigma_\varepsilon^2 I_{J\times J})$.

Theorem 1 indicates, consistent with the cited related research, that the asymptotic properties of the intricate WALS estimator g is governed by the asymptotic properties of the more straightforward weighted estimator $\tilde{\phi} = \eta(\hat{\phi})\hat{\phi}$ of ϕ, the mean vector of a normally distributed random vector with scale parameter σ_ε^2. That is, the optimal solution for estimating ϕ, using the weighted estimator $\tilde{\phi}$ is also best for estimating γ, employing the weighted estimator g, in any classic linear regression in the face of uncertain exact linear restrictions. Particularly, the preferred weight function for the WALS estimator g is determined by the "best" weight function for forming $\tilde{\phi} = \eta(\hat{\phi})\hat{\phi}$. Note that as σ_ε^2 can be consistently estimated using the OLS residuals, Theorem 1 holds when the error variance is unknown and estimated using the usual OLS estimator.

The findings of Theorem 1 explicitly require that the regressors are such that $plimA = plim(W'\varepsilon/n) = 0_{k\times1}$, as has the research before us. We now extend Theorem 1 to allow for regressors that may be correlated with the error term. Explicitly, we assume $plimA = plim(W'\varepsilon/n) = \rho$, and for reasons that will be clear in the next section, we also assume that $\rho = \rho_0/\sqrt{n}$, with ρ_0 fixed; i.e., that ρ is in a $\sqrt{n}$ −shrinking neighbourhood of zero. This is a commonly adopted assumption when exploring the asymptotic distribution of a test of whether $plim(W'\varepsilon/n) = 0_{k\times1}$ (e.g., see Hausman, 1978; Morey, 1984) and our use of this assumption here means we are considering values of ρ close to this belief. In this sense, our results are restrictive, but we hope they still provide guidance for a wider range of ρ values[14].

Under this framework, it follows that:

$$\hat{\phi} = \left[RD^{*-1}R'\right]^{-1/2}R\psi + \left[RD^{*-1}R'\right]^{-1/2}RD^{*-1}\rho_0 + \left[RD^{*-1}R'\right]^{-1/2}RD^{*-1}V_1^*,$$

$$\sqrt{n}(g_u - \gamma) = D^{*-1}\rho_0 + D^{*-1}V_1^*,$$

$$\sqrt{n}(g_r - \gamma) = -D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\psi + \left[I_{k\times k} - D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\right]D^{*-1}\rho_0 + \left[I_{k\times k} - D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\right]D^{*-1}V_1^*,$$

$$\sqrt{n}(g_u - g_r) = D^{*-1}R'\left[RD^{*-1}R'\right]^{-1}R\left(D^{*-1}V_1^* + \psi + D^{*-1}\rho_0\right)$$

where $V_1^* = \sqrt{n}A^*$ with $A^* = (W'\varepsilon/n - \rho)$, and $D^* = D/n$. As before, let $M_{WW} = plim\ D^*$, $F = M_{WW}^{-1}R'[RM_{WW}^{-1}R']^{-\frac{1}{2}}$, $\Psi = M_{WW}^{-1} - FF'$, $\phi = [RM_{WW}^{-1}R']^{-\frac{1}{2}}R\psi$ and, also, let $\phi^* = \phi + F'\rho_0$. Then, noting that $V_1^* \overset{d}{\to} U_1^* \sim N(0, \sigma_\varepsilon^2 M_{WW})$, we have:

$$\begin{pmatrix} \hat{\phi} \\ \sqrt{n}(g_r - \gamma) \\ \sqrt{n}(g_u - \gamma) \\ \sqrt{n}(g_u - g_r) \end{pmatrix} \overset{d}{\to} U^* \sim N(\kappa^*, \Sigma) \tag{15}$$

where $U^* = \begin{pmatrix} U_2^* \\ U_3^* \\ U_4^* \\ U_5^* \end{pmatrix} = \begin{pmatrix} \phi^* + F'U_1^* \\ -F\phi^* + M_{WW}^{-1}\rho_0 + \Psi U_1^* \\ M_{WW}^{-1}\rho_0 + M_{WW}^{-1}U_1^* \\ FU_2^* \end{pmatrix}$, $\kappa^* = \begin{pmatrix} \phi^* \\ -F\phi^* + M_{WW}^{-1}\rho_0 \\ M_{WW}^{-1}\rho_0 \\ F\phi^* \end{pmatrix}$ and

$$\Sigma = \sigma_\varepsilon^2 \begin{pmatrix} I_{J\times J} & 0_{J\times k} & F' & F' \\ 0_{k\times J} & \Psi & \Psi & 0_{k\times k} \\ F & \Psi & M_{WW}^{-1} & FF' \\ F & 0_{k\times k} & FF' & FF' \end{pmatrix}.$$

The asymptotic bias function for g is then:

$$AB(g,\gamma) = E(\eta(\phi^*)(FU_2^*) + U_3^*)$$
$$= FE(\eta(\phi^*)U_2^* - \phi^*) + M_{WW}^{-1}\rho_0.$$

As $L_1(g,\gamma) \overset{d}{\to} \|\eta(\phi^*)(FU_2^*) + U_3^*\|_\Pi^2$ and

$$L_2(g,\gamma) \overset{d}{\to} (\eta(\phi^*)(FU_2^*) + U_3^*)(\eta(\phi^*)(FU_2^*) + U_3^*)',$$

[14] Whether exploring the properties of estimators under a local misspecification framework is reasonable has been questioned (see, e.g., Raftery and Zheng, 2003; Hjort and Claeskens, 2003b), but is regularly adopted to explore large sample properties.

we have the following asymptotic risk functions:

$$AR_1(g,\gamma) = E\|\eta(\phi^*)(FU_2^*) + U_3^*\|_\Pi^2$$

$$= E\left\|F(\eta(\phi^*)U_2^* - \phi^*) + \left(U_3^* - E(U_3^*)\right) + M_{WW}^{-1}\rho_0\right\|_\Pi^2$$

$$= E\|\eta(\phi^*)(FU_2^*) - F\phi^*\|_\Pi^2 + E\|U_3^* - E(U_3^*)\|_\Pi^2 + \rho_0' M_{WW}^{-1}\Pi M_{WW}^{-1}\rho_0$$

$$+2\rho_0' M_{WW}^{-1}\Pi F E(\eta(\phi^*)U_2^* - \phi^*)$$

$$= E\|\eta(\phi^*)U_2^* - \phi^*\|_{\mathrm{F'\Pi F}}^2 + \sigma_\varepsilon^2 tr(\Pi\Psi) + \rho_0' M_{WW}^{-1}\Pi M_{WW}^{-1}\rho_0$$

$$+2\rho_0' M_{WW}^{-1}\Pi F\left(E(\eta(\phi^*)U_2^* - \phi^*)\right)$$

and

$$AR_2(g,\gamma) = E\left((\eta(\phi^*)(FU_2^*) + U_3^*)(\eta(\phi^*)(FU_2^*) + U_3^*)'\right)$$

$$= E\left((\eta(\phi^*)(FU_2^*) - F\phi^* + U_3^* - E(U_3^*) + M_{WW}^{-1}\rho_0)(\eta(\phi^*)(FU_2^*) - F\phi^* + U_3^* - E(U_3^*) + M_{WW}^{-1}\rho_0)'\right)$$

$$= E\left((\eta(\phi^*)(FU_2^*) - F\phi^*)(\eta(\phi^*)(FU_2^*) - F\phi^*)'\right)$$

$$+E(U_3^* - E(U_3^*))(U_3^* - E(U_3^*))') + M_{WW}^{-1}\rho_0\rho_0' M_{WW}^{-1}$$

$$+M_{WW}^{-1}\rho_0 E(\eta(\phi^*)U_2^* - \phi^*)'F' + F\left(E(\eta(\phi^*)U_2^* - \phi^*)\right)\rho_0' M_{WW}^{-1}$$

$$= F\left(E((\eta(\phi^*)U_2^* - \phi^*)(\eta(\phi^*)U_2^* - \phi^*)')\right)F' + \sigma_\varepsilon^2\Psi + M_{WW}^{-1}\rho_0\rho_0' M_{WW}^{-1}$$

$$+M_{WW}^{-1}\rho_0(E(\eta(\phi^*)U_2^* - \phi^*)')F' + F\left(E(\eta(\phi^*)U_2^* - \phi^*)\right)\rho_0' M_{WW}^{-1}.$$

Theorem 2, given below, follows from utilizing these results.

Theorem 2: Under the framework of this section, with $plim(W'\varepsilon/n) = \rho = \rho_0/\sqrt{n}$, fixed, the asymptotic bias, variance and risk functions of the WALS estimator are:

$$AB(g,\gamma) = F[AB(\eta(\phi^*)U_2^*, \phi^*)] + M_{WW}^{-1}\rho_0,$$

$$AV(g,\gamma) = F[AV(\eta(\phi^*)U_2^*)]F' + \sigma_\varepsilon^2\Psi,$$

$$AR_1(g,\gamma) = E\|\eta(\phi^*)U_2^* - \phi^*\|_{\mathrm{F'\Pi F}}^2 + \sigma_\varepsilon^2 tr(\Pi\Psi) + \rho_0' M_{WW}^{-1}\Pi M_{WW}^{-1}\rho_0$$

$$+2\rho_0' M_{WW}^{-1}\Pi F\left(E(\eta(\phi^*)U_2^* - \phi^*)\right),$$

and

$$AR_2(g,\gamma) = F[AR_2(\eta(\phi^*)U_2^*, \phi^*)]F' + \sigma_\varepsilon^2\Psi + M_{WW}^{-1}\rho_0\rho_0' M_{WW}^{-1}$$

$$+M_{WW}^{-1}\rho_0[AB(\eta(\phi^*)U_2^*, \phi^*)]'F' + F[AB(\eta(\phi^*)U_2^*, \phi^*)]\rho_0' M_{WW}^{-1},$$

where the $(J \times 1)$ vector U_2^* is distributed as $U_2^* \sim N(\phi^*, \sigma_\varepsilon^2 I_{J\times J})$, with $\phi^* = \phi + F'\rho_0$.

Theorem 2 shows that the asymptotic risk of the WALS estimator g, even when the regressors are potentially correlated with the error term, is determined by the asymptotic properties of the weighted estimator $\tilde{\phi} = \eta(\hat{\phi})\hat{\phi}$ of $\phi^* = \phi + F'\rho_0$, the mean vector of a normally distributed random vector with scale parameter σ_ε^2. That is, for a given situation for the regression model and correlation between the regressors and the disturbance term, the asymptotic risk function of the WALS estimator g is only determined by the properties of the weighted estimator $\tilde{\phi} = \eta(\hat{\phi})\hat{\phi}$ as an estimator of ϕ^*. We note, however, that the asymptotic risk for the WALS estimator g does depend on both the asymptotic risk and bias of $\tilde{\phi}$ when $\rho_0 \neq 0$.

Especially, the theorem implies that the "best" weight function for the WALS estimator g is given by the "best" weight function for deciding on $\tilde{\phi} = \eta(\hat{\phi})\hat{\phi}$. In terms of our assumption that σ_ε^2 is known, it is straightforward to show that using OLS residuals to form an estimator of σ_ε^2 would not lead to a consistent estimator when $\rho \neq 0$, but rather the usual estimator would converge to, say, σ_ε^{*2} $(= \sigma_\varepsilon^2 + \rho' M_{WW}^{-1}\rho$ under our assumptions); this effect would not alter the qualitative form of the optimal weight function.

In the next section we apply Theorem 2 to the case of weighting the OLS and 2SLS estimators for the setup we outlined at the beginning of this paper.

4. Applying Theorem 2 to Form a Weighted OLS and 2SLS Estimator

Returning to the setup outlined in Section 1, recall that the structural equation being estimated is $y = X_1\beta_1 + X_2\beta_2 + u = X\beta + u$, $X_2 = \mu + v$, where there are K_1 exogenous regressors in X_1 and potentially K_2 endogenous explanatory variables. The OLS and 2SLS estimators of β, $\hat{\beta}_{(1)}$ and $\hat{\beta}_{(2)}$ respectively, are defined in equation (2). We are interested in the asymptotic bias and asymptotic risk functions of the WALS estimator $\hat{\beta}_W$, which is a weighted combination of $\hat{\beta}_{(1)}$ and $\hat{\beta}_{(2)}$. We show that this question can be addressed by applying Theorem 2 from the previous section, implying the same choice of the asymptotically optimal weight function.

We proceed by examining an auxiliary regression of X_2 on Z. Let $\hat{X}_2$ be the matrix of fitted values from this regression and e be the associated matrix of residuals; i.e., $\hat{X}_2 = Z(Z'Z)^{-1}Z'X_2$ and $e = (I_{n\times n} - Z(Z'Z)^{-1}Z')X_2$. Consider the model

$$y = X_1\beta_1 + \hat{X}_2\beta_2 + e\alpha + u, \tag{16}$$

which we write as

$$y = X^*\beta^* + u, \tag{17}$$

with $X^* = [X_1 \quad \hat{X}_2 \quad e] = [X_s \quad e]$ and $\beta^* = [\beta_1' \quad \beta_2' \quad \alpha']' = [\beta_s' \quad \alpha']'$; there are $K^* = K_1 + 2K_2$ coefficients to be estimated. Let $\beta_u^* = [\beta_{s,u}' \quad \alpha_u']'$ be the OLS estimator of β^*. As e is orthogonal to both X_1 and $\hat{X}_2$, it can be readily shown that the OLS estimator of β_s^*, $\beta_{s,u}$, is identical to the

2SLS estimator, $\hat{\beta}_{(2)}$, for the structural model (1); e.g., see Hausman (1978), Nakamura and Nakamura (1981), Morey (1984). Further, $\alpha_u = (e'e)^{-1}e'y$. That is, OLS applied to (16) results in

$$\beta_u^* = \begin{bmatrix} \beta_{s,u} \\ \alpha_u \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{(2)} \\ \alpha_u \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{(2,1)} \\ \hat{\beta}_{(2,2)} \\ \alpha_u \end{bmatrix} = \begin{bmatrix} (X_1'C_1X_1)^{-1}X_1'C_1y \\ (X_2'C_2X_2)^{-1}X_2'C_2y \\ (e'e)^{-1}e'y \end{bmatrix}$$

where $C_1 = I_{n\times n} - X_2(X_2'P_2X_2)^{-1}X_2'P_2$, $C_2 = P_2 - P_3$ and $P_3 = X_1(X_1'X_1)^{-1}X_1{}'$. Given that

$$plim(X^{*\prime}u/n) = \begin{bmatrix} plim(X_1'u/n) \\ plim(\hat{X}_2'u/n) \\ plim(e'u/n) \end{bmatrix} = \begin{bmatrix} 0_{K_1\times 1} \\ 0_{K_2\times 1} \\ \delta \end{bmatrix} = \omega$$

it follows that

$$plim(\beta_u^*) = \begin{bmatrix} \beta_{s,u} \\ \alpha_u \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{(2)} \\ \alpha_u \end{bmatrix} = \begin{bmatrix} plim(\hat{\beta}_{(2,1)}) \\ plim(\hat{\beta}_{(2,2)}) \\ plim(\alpha_u) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_2 + B_1^{-1}\delta \end{bmatrix}$$

where $B_1 = Q_{X_2X_2} - Q_{X_2Z}Q_{ZZ}^{-1}Q_{ZX_2}$. Note, in particular, that $plim(\alpha_u) = \beta_2$ when $\delta = 0$.

Now consider the K_2 restrictions $\beta_2 = \alpha$ on model (16), which can be written in the form $R\beta^* = r$, with the $(K_2 \times K^*)$ matrix $R = [0_{K_2\times K_1} \quad I_{K_2\times K_2} \quad -I_{K_2\times K_2}]$ and the K_2-dimensional vector $r = 0_{K_2\times 1}$; obviously, this is also equivalent to the restrictions $\delta = 0_{K_2\times 1}$. Imposing that $\beta_2 = \alpha$ on model (16), the resulting RLS estimator of β^*, say $\beta_r^* = [\beta_{s,r}' \quad \alpha_r']'$, is the OLS estimator of β for the structural equation (1),

$$\beta_r^* = \begin{bmatrix} \beta_{s,r} \\ \alpha_r \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{(1)} \\ \hat{\beta}_{(1,2)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{(1,1)} \\ \hat{\beta}_{(1,2)} \\ \hat{\beta}_{(1,2)} \end{bmatrix} = \begin{bmatrix} (X_1'C_3X_1)^{-1}X_1'C_3y \\ (X_2'C_4X_2)^{-1}X_2'C_4y \\ (X_2'C_4X_2)^{-1}X_2'C_4y \end{bmatrix}$$

where, with $P_4 = X_2(X_2'X_2)^{-1}X_2{}'$, $C_3 = I_{n\times n} - P_4$ and $C_4 = I_{n\times n} - P_3$. As a consequence,

$$plim(\beta_r^*) = \begin{bmatrix} plim(\hat{\beta}_{(1,1)}) \\ plim(\hat{\beta}_{(1,2)}) \\ plim(\hat{\beta}_{(1,2)}) \end{bmatrix} = \begin{bmatrix} \beta_1 - B_2^{-1}B_3\delta \\ \beta_2 + B_4^{-1}\delta \\ \beta_2 + B_4^{-1}\delta \end{bmatrix}$$

where $B_2 = Q_{X_1X_1} - Q_{X_1X_2}Q_{X_2X_2}^{-1}Q_{X_2X_1}$, $B_3 = Q_{X_1X_2}Q_{X_2X_2}^{-1}$ and $B_4 = Q_{X_2X_2} - Q_{X_2X_1}Q_{X_1X_1}^{-1}Q_{X_1X_2}$. That is, whereas both β_u^* and β_r^* are inconsistent estimators of β^*, $\beta_{s,u} = \hat{\beta}_{(2)}$, the 2SLS estimator, is consistent for β.

In summary, the OLS and 2SLS estimators of β for the structural model (1) are, respectively, RLS and OLS estimators of β_s for the linear regression model (16) with restrictions $\beta_2 = \alpha$, which is identical to imposing that $\delta = 0$. The weighted estimator of β^* for model (16) or (17) is:

$$\beta_W^* = \lambda(\hat{\phi})\beta_u^* + \left(1 - \lambda(\hat{\phi})\right)\beta_r^* \tag{18}$$

where $\lambda(.)$ is a continuous, random function, dependent on $\hat{\phi} = \left[R\left(X^{*\prime}X^*\right)^{-1}R'\right]^{-\frac{1}{2}}(R\beta_u^* - r)$, a measure of how close the K_2 restrictions $\beta_2 = \alpha$ hold in the sample. Note that $\beta_W^* = [\hat{\beta}_W' \quad \hat{\alpha}_W']'$, where $\hat{\beta}_W$ is the estimator of interest that weights the OLS and 2SLS estimators of β. The weight function, $\lambda(\hat{\phi})$, likely also depends on an estimator of σ^2; we assume that σ^2 is known but, unlike the general case considered for Theorem 2, here a consistent estimator of σ^2 is available. Clearly, we have a particular application of the framework covered by Theorem 2.

Specifically, for the problem at hand, with $\sqrt{n}(\beta_W^* - \beta^*) = \lambda(\hat{\phi})\left(\sqrt{n}(\beta_u^* - \beta_r^*)\right) + \sqrt{n}(\beta_r^* - \beta^*)$, some algebraic manipulations show that:

(i) $\hat{\phi} = \left[\left(\hat{X}_2' C_4 \hat{X}_2\right)^{-1} + (e'e)^{-1}\right]^{-\frac{1}{2}}(\hat{\beta}_{(2,2)} - \alpha_u)$;

(ii) $M_{X^*X^*} = plim\left(X^{*\prime}X^*/n\right) = \begin{bmatrix} Q_{X_1X_1} & Q_{X_1X_2} & 0_{K_1\times K_2} \\ Q_{X_2X_1} & Q_{X_2X_2} - B_1 & 0_{K_2\times K_2} \\ 0_{K_2\times K_1} & 0_{K_2\times K_2} & B_1 \end{bmatrix} = \begin{bmatrix} Q_s & 0_{K\times K_2} \\ 0_{K_2\times K} & B_1 \end{bmatrix}$

where the $(K \times K)$ matrix $Q_s = \begin{bmatrix} Q_{X_1X_1} & Q_{X_1X_2} \\ Q_{X_2X_1} & Q_{X_2X_2} - B_1 \end{bmatrix}$;

(iii) $\left[RM_{X^*X^*}^{-1}R'\right] = (B_4 - B_1)^{-1} + B_1^{-1}$;

(iv) $RM_{X^*X^*}^{-1}\omega = -B_1^{-1}\delta$;

(v) $(\beta^* - \beta_0^*) = \left[0_{K_1\times 1}\ 0_{K_2\times 1}\ \alpha - \beta_2\right]$, where β_0^* is a solution to $\beta_2 = \alpha$.

Given the equivalence of imposing the restrictions $\beta_2 = \alpha$ or $\delta = 0$ on model (16), we suppose that $\beta_2 - \alpha = G\delta$, where G is a nonstochastic matrix; e.g., one possibility is $G = -B_1^{-1}$. Then, with $F^* = M_{X^*X^*}^{-1}R'\left[RM_{X^*X^*}^{-1}R'\right]^{-\frac{1}{2}}$ and assuming, as before, that it is reasonable to let $\delta = \delta_0/\sqrt{n}$ to give $\omega = \omega_0/\sqrt{n}$, further algebra reveals that:

(vi) $F^{*\prime}\omega_0 = -[(B_4 - B_1)^{-1} + B_1^{-1}]^{-\frac{1}{2}}B_1^{-1}\delta_0$;

(vii) $\phi = \left[RM_{X^*X^*}^{-1}R'\right]^{-\frac{1}{2}}G\delta_0 = [(B_4 - B_1)^{-1} + B_1^{-1}]^{-\frac{1}{2}}G\delta_0$;

(viii) $\phi^* = [(B_4 - B_1)^{-1} + B_1^{-1}]^{-\frac{1}{2}}(G - B_1^{-1})\delta_0$.

To apply Theorem 2, let $AB(\beta_W^*, \beta^*), AV(\beta_W^*, \beta^*), AR_1(\beta_W^*, \beta^*)$ and $AR_2(\beta_W^*, \beta^*)$ denote the asymptotic bias, variance and risk functions of β_W^* as an estimator of β^*. The loss functions for the risk functions are, respectively, $L_1(\beta_W^*, \beta^*) = \left\|\sqrt{n}(\beta_W^* - \beta^*)\right\|_{\Pi_n^*}^2$ and the matrix loss function $L_2(\beta_W^*, \beta^*) = (n(\beta_W^* - \beta^*)(\beta_W^* - \beta^*)')$, where, here, Π_n^* is a random, positive semi-definite $(K^* \times K^*)$ matrix. We shortly discuss possible forms of Π_n^*. Then, with $\Psi^* = M_{X^*X^*}^{-1} - F^*F^{*\prime}$ and assuming that $\frac{1}{n}\Pi_n^* \xrightarrow{p} \Pi^*$, applying Theorem 2 leads to Corollary 1 below.

Corollary 1: Under the framework of this section, the asymptotic bias, variance and risk functions of the WALS estimator, β_W^*, are:

$$AB(\beta_W^*, \beta^*) = F^*[AB(\lambda(\phi^*)U_2^*, \phi^*)] + M_{X^*X^*}^{-1}\omega_0,$$

$$AV(\beta_W^*, \beta^*) = F^*[AV(\lambda(\phi^*)U_2^*)]F^{*\prime} + \sigma^2\Psi^*,$$

$$AR_1(\beta_W^*, \beta^*) = E\|\lambda(\phi^*)U_2^* - \phi^*\|_{F^{*\prime}\Pi^*F^*}^2 + \sigma^2 tr(\Pi^*\Psi^*) + \omega_0' M_{X^*X^*}^{-1}\Pi^* M_{X^*X^*}^{-1}\omega_0$$
$$+2\omega_0' M_{X^*X^*}^{-1}\Pi^* F^*\big(E(\lambda(\phi^*)U_2^* - \phi^*)\big),$$

and

$$AR_2(\beta_W^*, \beta^*) = F^*[AR_2(\lambda(\phi^*)U_2^*, \phi^*)]F^{*\prime} + \sigma^2\Psi^* + M_{X^*X^*}^{-1}\omega_0\omega_0' M_{X^*X^*}^{-1}$$
$$+M_{X^*X^*}^{-1}\omega_0[AB(\lambda(\phi^*)U_2^*, \phi^*)]' F^{*\prime} + F^*[AB(\lambda(\phi^*)U_2^*, \phi^*)]\omega_0' M_{X^*X^*}^{-1},$$

where the $(K_2 \times 1)$ vector U_2^* is distributed as $U_2^* \sim N(\phi^*, \sigma^2 I_{K_2\times K_2})$.

These outcomes show that the asymptotic properties of the complex weighted estimator β_W^* of β^*, which includes as a subset the coefficient vector of fundamental interest β, are determined by the asymptotic features of the far simpler weighted estimator $\tilde{\phi} = \lambda(\hat{\phi})\hat{\phi}$ as an estimator of ϕ^*, the mean vector of a normally distributed random vector.

In practice, our fundamental interest lies in weighting $\hat{\beta}_{(2)} = \beta_{s,u}$ and $\hat{\beta}_{(1)} = \beta_{s,r}$ using $\lambda(\hat{\phi})$, to form the weighted estimator $\hat{\beta}_W$, the sub-vector of β_W^* consisting of its first K elements:

$$\hat{\beta}_W = \lambda(\hat{\phi})\beta_{s,u} + \left(1 - \lambda(\hat{\phi})\right)\beta_{s,r} = \lambda(\hat{\phi})\hat{\beta}_{(2)} + \left(1 - \lambda(\hat{\phi})\right)\hat{\beta}_{(1)}. \quad (19)$$

To apply Corollary 1 to obtain the asymptotic properties of $\hat{\beta}_W$, we employ the following:

(i) S is a $(K \times K^*)$ selector matrix with $S = [I_{K\times K} \quad 0_{K\times K_2}]$;
(ii) $\hat{\beta}_W = S\beta_W^*$;
(iii) $AB(\hat{\beta}_W, \beta) = S(AB(\beta_W^*, \beta^*))$;
(iv) $M_{X^*X^*}^{-1}\omega_0 = [0_{1\times K} \quad \delta_0' B_1^{-1}]'$;
(v) $SM_{X^*X^*}^{-1}\omega_0 = 0_{K\times 1}$;
(vi) $AV(\hat{\beta}_W, \beta) = S(AV(\beta_W^*, \beta^*))S'$.

Then, let Π_n^* be appropriately partitioned as $\Pi_n^* = \begin{bmatrix}\Pi_{11n}^* & \Pi_{12n}^* \\ \Pi_{21n}^* & \Pi_{22n}^*\end{bmatrix}$. As traditional choices of this weight matrix are I_n and $X^{*\prime}X^*$, both of which lead to $\Pi_{12n}^* = 0_{K\times K_2}$ and $\Pi_{21n}^* = 0_{K_2\times K}$, for our case, we adopt the assumption that $\Pi_n^* = \begin{bmatrix}\Pi_{11n}^* & 0_{K\times K_2} \\ 0_{K_2\times K} & \Pi_{22n}^*\end{bmatrix}$. Applying this additional assumption with the provided results, gives rise to:

(vii) $S'S\Pi_n^*S'S = \Pi_{11n}^{**}$, with $\Pi_{11n}^{**} = \begin{bmatrix}\Pi_{11n}^* & 0_{K\times K_2} \\ 0_{K_2\times K} & 0_{K_2\times K_2}\end{bmatrix}$ and $\frac{1}{n}\Pi_{11n}^{**} \xrightarrow{p} \Pi_{11}^{**}$;

(viii) $\Pi_{11}^{**} M_{X^*X^*}^{-1} \omega_0 = 0_{(K+K_2)\times 1}$;

(ix) $L_1(\beta_W^*, \beta^*) = \|\hat{\beta}_W - \beta\|_{\Pi_{11n}^*}^2 + \|\hat{\alpha}_W - \alpha\|_{\Pi_{22n}^*}^2 = L_1(\hat{\beta}_W, \beta) + L_1(\hat{\alpha}_W, \alpha)$;

(x) $L_1(\hat{\beta}_W, \beta) = \|\beta_W^* - \beta^*\|_{\Pi_{11n}^{**}}^2$;

(xi) $AR_1(\beta_W^*, \beta^*) = AR_1(\hat{\beta}_W, \beta) + AR_1(\hat{\alpha}_W, \alpha)$;

(xii) $AR_1(\hat{\beta}_W, \beta) = E\|\lambda(\phi^*)(FU_2^*) + U_3^*\|_{\Pi_{11}^{**}}^2$;

(xiii) $L_2(\hat{\beta}_W, \beta) = S\{L_2(\beta_W^*, \beta^*)\}S'$;

(xiv) $AR_2(\hat{\beta}_W, \beta) = S\{AR_2(\beta_W^*, \beta^*)\}S'$.

Employing Corollary 1, with (i) to (xiv), generates Corollary 2, which gives the asymptotic properties of the weighted estimator $\hat{\beta}_W$:

Corollary 2: Under the framework of this section, the asymptotic bias, variance and risk functions of the WALS estimator, $\hat{\beta}_W$, are:

$$AB(\hat{\beta}_W, \beta) = S\{F^*[AB(\lambda(\phi^*)U_2^*, \phi^*)]\},$$

$$AV(\hat{\beta}_W, \beta) = S\{F^*[AV(\lambda(\phi^*)U_2^*)]F^{*\prime}\}S' + \sigma^2 S\Psi^* S',$$

$$AR_1(\hat{\beta}_W, \beta) = E\|\lambda(\phi^*)U_2^* - \phi^*\|_{F^{*\prime}\Pi_{11}^{**}F^*}^2 + \sigma^2 tr(\Pi_{11}^{**}\Psi^*),$$

and

$$AR_2(\hat{\beta}_W, \beta) = S\{F^*[AR_2(\lambda(\phi^*)U_2^*, \phi^*)]F^{*\prime}\}S' + \sigma^2 S\Psi^* S',$$

where the $(K_2 \times 1)$ vector U_2^* is distributed as $U_2^* \sim N(\phi^*, \sigma^2 I_{K_2\times K_2})$.

These asymptotic properties, provided in Corollary 2, are of identical form to those given under Theorem 1, despite the potential endogeneity of some of the explanatory factors. How best to compose the WALS estimator of β, $\hat{\beta}_W$, which combines the OLS and 2SLS estimators of β, is governed solely by determining the preferred method of constructing the estimator $\tilde{\phi} = \lambda(\hat{\phi})\hat{\phi}$ for ϕ^*, the mean vector of a normally distributed variate. Obviously, approaches concluded to be favoured in the original WALS literature might also be useful when applying Corollary 2. We now turn to this question for our empirical applications, drawing from some of the existing literature on this issue.

5. Weight Functions and Empirical Examples

Magnus, along with his many co-authors, advocate a Bayesian type weighting scheme to form the WALS estimator in their analyses of combining "focus" and "auxillary" variables, giving rise to a WALS estimator that is a mixture of frequentist and Bayesian concepts; the estimator averages estimates from frequentist restricted and ordinary least squares with weights formed using Bayesian ideas. Philosophically, it might be viewed as contentious to form such hybrid estimators, an issue that we do not pursue here. Our focus is on combining information from two regressions in a manner suggested in other contexts to provide reasonable weights. To proceed, we note that these methods form "preferred" (in some sense) estimators, $\tilde{\phi}$, of ϕ^*. We then calculate an estimate of the weight λ, using that $\tilde{\phi} = \lambda\hat{\phi}$; in the vector case we have $\hat{\lambda} = \left(\frac{\tilde{\phi}'\tilde{\phi}}{\hat{\phi}'\hat{\phi}}\right)^{1/2}$. The model averaging estimator $\hat{\beta}_W$ can then be easily formed. Here we provide only salient details of the approach advocated by this WALS body of research, referring the reader to, for instance, Magnus et al. (2010), Kumar and Magnus (2013), and Magnus and De Luca (2016) for details and discussions.

As we are estimating the mean vector, ϕ^*, of i.i.d. normally distributed random variates, the focus of this aforementioned research is on how to estimate one element of the vector, $\phi_j^*, j = 1, \dots, K_2$. Let x be a univariate random variate with $x|\tau \sim N(\tau, 1)$, with prior density $\pi(\tau)$ and subsequent posterior density $p(\tau|x)$. The mean and variance of this posterior density, say $\hat{\tau}$ and $var(\hat{\tau})$ are used as desired estimators. For various reasons (as detailed by the authors), Magnus et al. advocate that the prior be chosen from the three-parameter reflected generalized gamma distribution with density:

$$\pi(\tau) = \frac{qc^{\Lambda}}{2\Gamma(\Lambda)}|\tau|^{-\vartheta}e^{-c|\tau|^q}$$

where $-\infty < \tau < \infty, q > 0, 0 \le \vartheta < 1, c > 0$ and $\Lambda = (1-\vartheta)/q$. Particular examples are the Laplace ($q = 1, \vartheta = 0, \Lambda = 1$), the reflected Weibull ($\Lambda = 1, \vartheta = 1 - q$) and the Subbotin ($\vartheta = 0, \Lambda = 1/q$, with $0 < q < 1$) densities. Aside from other features, these choices of prior can be chosen to be neutral, in the sense that there is ignorance about whether or not τ is positive or negative, and whether $|\tau|$ is larger or smaller than one[15]. This latter notion of ignorance, adopted by this body of work, is framed in terms of a classical regression model with one possible auxiliary regressor (say W_2, in terms of the notation in section 2) whose coefficient, γ_2, may or may not be zero. The adjusted R^2 ($\bar{R}^2$) increases with the inclusion of the auxiliary variable W_2 iff the relevant t-ratio exceeds one in absolute value, providing one motivation for defining ignorance in terms of $|\tau| >< 1$. In addition, a mean squared error (MSE) comparison of the RLS and OLS estimators (with and without the auxiliary variable) shows that the MSE of the RLS estimator is lower than that of the OLS estimator when (twice) the non-centrality parameter (say NP), associated with the t- or F-test of $\beta_2 = 0$, is less than 1, and the MSEs are equal when $2NP = 1$. This is a well-noted result in the literature; see, e.g., Magnus and Durbin (1999, Theorem 1), who mention this in terms of the t-ratio and cite earlier research. Toro-Vizcarrondo and Wallace (1968) generalize this outcome to J exact linear restrictions (of the from $R\gamma = r$ in terms of Section 3) within the classical linear regression framework. Specifically, Toro-Vizcarrondo and Wallace show that the difference between the MSE matrices of the RLS and OLS estimators is positive semi-definite when (twice) the non-centrality parameter (NP) associated with the classical F-test of the validity of the J exact linear restrictions is less than 1. This implies, in terms of ϕ as given in Section 3, that the difference in MSE matrices is positive semi-definite when $\phi'\phi/\sigma_\varepsilon^2 \le 1$. That is, for our research, it remains likely reasonable to continue to think about ignorance in terms of whether $|\tau|$ is larger or smaller than one.

---

[15] That is, ignorance about the correlation between the regressor and the error term and the sign of any such correlation should it exist.

Turning to the question of prior distributions, in their study of mixing "focus" and "auxiliary" variables, Magnus et al. (2010) explore use of a Laplace prior with $c = log2$ (for neutrality) based on recommendations from Magnus (2002), who shows that $\hat{\tau}$ is admissible with bounded risk and desirable properties around $|\tau| = 1$. The Subbotin prior is the focus of Kumar and Magnus' (2013) research, with their work showing that the Laplace prior is not (Bayesian) robust whereas the Subbotin prior satisfies this property, as well as having lower minimax regret. Calculating c to achieve neutrality with a search of q over the range 0 and 1, the authors find that maximum regret is lowest for the Subbotin prior when $q = 0.7995$ and $c = 0.9377$. Accordingly, we use this prior with these q and c values in our empirical applications. As an earlier working paper version of Kumar and Magnus' paper (Einmahl et al., 2011) recommends the Subbotin prior with $q = 0.5$ in practical applications (with $c = 1.6783$ for neutrality), we additionally consider this prior. The minimax regret results of Kumar and Magnus observe that the reflected Weibull prior (with $q = 0.8876, c = log2$) leads to a marginally smaller minimax regret than the other examined priors (including the Laplace and Subbotin priors). As Magnus and De Luca (2016) also advocate use of the reflected Weibull prior, we also use this prior in our applications. Employing a range of priors allows us to provide some indication of how outcomes differ across them (at least for our examples).

Analytic formulae for the mean and variance of the posterior distribution from using the Laplace prior are provided by Magnus et al. (2010, p145). Numerical integration routines are required when employing the reflected Weibull and Subbotin priors, with Magnus and De Luca (2016), for instance, giving details on calculating the mean and variance of the posterior distribution. We undertook our empirical work using Stata /SE version 14.2 (StataCorp, 2015) utilizing the Stata module "integrate", provided by Mander (2012), to undertake the needed integrations. We now present results from our two illustrations.

5.1. Example 1: Returns to Schooling

Our first application of weighting OLS and 2SLS uses Example 15.4 from the popular undergraduate textbook Wooldridge (2016, pp 473-474), which is based on a subsample of data and one model from Card (1995)[16]. Card uses data from the National Longitudinal Survey of Young Men (NLSYM), which began in 1966 with 5525 men aged 14-24 and continued with follow-up surveys through 1981, with the subsample of interest consisting of 3010 men who provide valid responses regarding education and wages from the 1976 interview. Although compared with a nationally representative sample, as noted by Card, the NLSYM oversamples men from the Southern region and those who are classified as Black, following Card and Wooldridge we ignore any implications of this sampling design.

[16] We obtained the data from Cengage Learning's website for Wooldridge's textbook; see http://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20bI&product_isbn_issn=9781305270107, last accessed 10 March 2017.

The focus of the analysis is to ascertain the link between education and earnings, taking account of the fact that to explore for such a relationship perhaps needs an exogenous determinant of schooling options. One reason suggested in the literature as to why education may be correlated with the unobserved component of earnings is so-called "ability bias", whereby an individual may have an innate trait of being able to earn higher wages for any education level. If education is correlated with the error term, then a suitable instrumental variable is needed. Card suggests that accessibility of a college, near to where the man resided, is a reasonable instrumental variable. Table 1 reports some descriptive statistics for the data we consider, with definitions for the variables given in Appendix 1. Our results accord with those reported by Card (1995).

Table 1: Descriptive statistics for some key variables used in Example 1

Wage in 1976: hourly earnings in cents		
	Minimum	100
	Maximum	2404
	Mean	577
	Std. Dev.	263
Age in 1976: years		
	Minimum	24
	Maximum	34
	Mean	28.1
	Std. Dev.	3.1
Experience in 1976: years		
	Minimum	0
	Maximum	23
	Mean	8.9
	Std. Dev.	4.1
Education in 1976: years		
	Minimum	1
	Maximum	18
	Mean	13.3
	Std. Dev.	2.7
Percent Black		23.4
Percent lived in SMSA 1966		65.0
Percent lived in SMSA 1976		71.3
Percent raised near a 4-year college		68.2

As stated, Card's hypothesis is that the presence of a 4-year (in particular) college near where the man resided is a valid instrumental variable for education. The men in the sample, ceteris paribus, who have ready access to such a college, achieve significantly higher levels of education and earnings. Perhaps, Card suggests, having a college locally available reduces the expense of undertaking post secondary education and/or increases beliefs about the benefits of higher education.

One of Card's models (Card, 1995, p32, column (2) in Table 2), which forms the basis of Example 15.4 in Wooldridge (2016), is:

$$lwage_i = \beta_{11} + \beta_{12} exper_i + \beta_{13} exper_i^2 + \beta_{14} black_i + \beta_{15} smsa_i \qquad (20)$$

$$+\beta_{16} smsa66_i + \beta_{17} south_i + \sum_{j=2}^{9} \beta_{1(6+j)} reg66j_i + \beta_{21} educ_i + u_i,$$
$$i = 1, \dots, 3010.$$

The analysis is interested in the estimated education coefficient, β_{21}, which is loosely referred to as the "rate of return to schooling", an interpretation fraught with issues, as well acknowledged by Card. After exploring conceivable reasons for the possible endogeneity of $educ$, as well as motivating why $nearc4$ might be a valid instrumental variable for completed education, Card reports a number of reduced form IV/2SLS equations, including for the model provided in equation (20)[17]. This example gives $K_1 = 15, K_2 = 1, K = 16, K^* = 17, n = 3010, L_2 = 1$. The OLS and 2SLS estimates, along with standard errors, are reported in columns (2) and (3) of Table 2[18], whereas the outcomes for the WALS cases are reported in columns (4) through (7).

Although not reported here, the estimated coefficient of *nearc4* in the regression of *educ* on *nearc4* and the other explanatory variables from equation (20), suggests that, on average, proximity to a college in 1966 did indeed lead to more education of approximately one third of a year, holding other factors constant. The effect is statistically significant. For space reasons, we do not report the appropriate auxiliary regression (16) for this example, with the results being available on request. However, we note, before examining the estimation results, that the DWH statistic for this sample is 1.168 with an $F(1,2993)$ p-value of 0.280, which would suggest that *educ* is not correlated with the error term.

The OLS estimate of the education effect implies a 7.5% benefit in wages for each additional year of education, holding other factors constant. The estimate from using 2SLS is approximately 43% higher, indicating that an additional year of education leads to a 13.2% increase in expected wages, ceteris paribus. The WALS estimates are similar across the four priors, leading to an estimated 10.5% to 11% increase in estimated wages with an additional year of schooling, with the weight factor ranging from 0.541 and 0.622, each choice of prior slightly weighing the 2SLS estimates more than the OLS estimates. Using college proximity as an instrument, either with 2SLS or WALS, increases the estimated gain from education by over 40%. The impact of education on wages varies little across the four WALS estimates. The standard errors associated with the 2SLS and WALS education estimates are relatively large compared with those from using OLS, but the WALS standard errors are at least 16% smaller than those from using 2SLS.

---

[17] We note that Card explores whether other explanatory variables, in addition to *educ*, might be correlated with the error term. In particular, when *educ* is correlated with the error term then so too is *exper*, given the construction of this variable. We do not pursue this in our illustration.

[18] The reported 2SLS standard errors incorporate the common degrees-of-freedom adjustment.

Table 2: Estimated hourly earnings equation

Explanatory Variable	OLS	2SLS	WALS Laplace $q=1$ $c=log2$	WALS Subbotin $q=0.5$ $c=1.6783$	WALS Subbotin $q=0.7995$ $c=0.9377$	WALS Weibull $q=0.8876$ $c=log2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	4.621^* (0.074)	3.666^* (0.925)	4.027^* (0.766)	4.105^* (0.744)	4.051^* (0.761)	4.052^* (0.763)
Exper	0.085^* (0.007)	0.108^* (0.024)	0.099^* (0.020)	0.098^* (0.019)	0.099^* (0.020)	0.099^* (0.020)
exper^2	-0.0023^* (0.0003)	-0.0023^* (0.0003)	-0.0023^* (0.0003)	-0.0023^* (0.0003)	-0.0023^* (0.0003)	-0.0023^* (0.0003)
Black	-0.199^* (0.018)	-0.147^* (0.054)	-0.167^* (0.046)	-0.171^* (0.045)	-0.168^* (0.046)	-0.168^* (0.046)
Smsa	0.136^* (0.020)	0.112^* (0.032)	0.121^* (0.029)	0.123^* (0.028)	0.122^* (0.029)	0.122^* (0.029)
South	-0.148^* (0.026)	-0.145^* (0.027)	-0.146^* (0.027)	-0.146^* (0.027)	-0.146^* (0.027)	-0.146^* (0.027)
smsa66	0.026^* (0.019)	0.019^* (0.022)	0.021 (0.021)	0.022 (0.021)	0.022 (0.021)	0.022 (0.021)
reg662	0.096^* (0.036)	0.101^* (0.038)	0.099^* (0.038)	0.099^* (0.038)	0.099^* (0.038)	0.099^* (0.038)
reg663	0.145^* (0.035)	0.148^* (0.037)	0.147^* (0.037)	0.147^* (0.037)	0.147^* (0.037)	0.147^* (0.037)
reg664	0.055 (0.042)	0.050 (0.044)	0.052 (0.044)	0.052 (0.044)	0.052 (0.044)	0.052 (0.044)
reg665	0.128^* (0.042)	0.146^* (0.047)	0.139^* (0.046)	0.138^* (0.046)	0.139^* (0.046)	0.139^* (0.046)
reg666	0.141^* (0.045)	0.163^* (0.052)	0.154^* (0.050)	0.153^* (0.050)	0.154^* (0.050)	0.154^* (0.049)
reg667	0.118^* (0.045)	0.135^* (0.049)	0.128^* (0.049)	0.127^* (0.049)	0.128^* (0.049)	0.128^* (0.049)
reg668	-0.056 (0.051)	-0.083 (0.059)	-0.073 (0.058)	-0.071 (0.057)	-0.072 (0.058)	-0.072 (0.058)
reg669	0.119^* (0.039)	0.108^* (0.042)	0.112^* (0.041)	0.113^* (0.041)	0.112^* (0.041)	0.112^* (0.041)
Educ	0.075^* (0.003)	0.132^* (0.055)	0.110^* (0.045)	0.105^* (0.044)	0.109^* (0.045)	0.109^* (0.046)
$\hat{\lambda}$	*n.a.*	*n.a.*	0.622	0.541	0.597	0.596

Notes: Standard errors are provided in parentheses. A superscript * indicates statistical significance, against an appropriate one-sided alternative hypothesis (as relevant) with, at least, an approximate 10% significance level.

We have used a * alongside estimates in Table 2 to denote statistical significance at the nominal 10% significance level, using a standard normal critical value. It remains to be explored whether a normality assumption reasonably approximates the distribution of the WAL estimator (even asymptotically), given that the estimator is a Bayesian combination of frequentist estimators. Accordingly, care is needed when interpreting the *s for the WALS cases. However, as previously discussed, we can consider whether the t-ratios are greater than one in magnitude as one way to ascertain whether a regressor is reasonably correlated with the dependent variable. Here, the absolute t-ratios are greater than one for all cases and variables, suggesting that inclusion of each explanatory variable is sensible, irrespective of estimation method.

Comparing the outcomes for all regressors across the four WALS cases, we see that the coefficient estimates are similar, only differing by at most 5%. The standard errors are comparable, with those obtained from the Subbotin prior with $q = 0.5$ sometimes being marginally smaller. At least for this illustration, the choice of prior does not lead to significantly different magnitudes of the estimated coefficients and standard errors. Turning finally to the estimated weights, we observe that $\hat{\lambda} > 0.5$, irrespective of prior, so that the WALS estimates are weighting 2SLS more than OLS, with the Laplace prior resulting in the highest $\hat{\lambda}$ and the Subbotin prior ($q = 0.5$) leading to the lowest $\hat{\lambda}$, but the difference between the highest and lowest $\hat{\lambda}$ is only 15%.

In summary, using WALS, at least for this illustration, offers a viable alternative when there is uncertainty regarding the possible endogeneity of one or more explanatory variables. The results demonstrate that the approach can be readily applied, often leading to significant gains in standard errors over just applying 2SLS. We now look at our second example.

5.2. Example 2: Role of Religion in Economic Growth

This demonstration of combining OLS and 2SLS using our results is based on the work of Barro and McCleary (2003) and Durlauf et al. (2012), amongst others, who explore the question of whether religion impacts aggregate economic growth. Barro and McCleary (hereafter BM) use survey data on religiosity to estimate the effect of church attendance and religious beliefs, specifically belief in heaven and in hell, on economic growth for a cross-country panel of 41 countries over three time periods (1965 to 1975, 1975 to 1985 and 1985 to 1995). Due to data limitations, some countries are only included in one or more of the periods leading to an unbalanced panel, and the determinants of growth, aside from those related to religion, are as advocated in the empirical framework of Barro and Sala-i-Martin (2004, Chapter 12).

BM hypothesize that church attendance is an input to the religion sector whereas believing in hell and/or heaven is an output. Their empirical results, based on instrumental variables estimation of their pooled data, suggest that economic growth correlates positively with religious beliefs, which BM interpret as consistent with a framework in which these beliefs lead to individual traits that result in increased productivity and hence economic growth. In contrast, church attendance is estimated to negatively correlate with economic growth, a feature that the authors suggest arises because higher church attendance indicates that more resources are being devoted to the religion sector, holding the output from the sector (religious beliefs) constant. In addition, BM postulate that church attendance might be representative of the impact that organized religion has on laws and regulations in the country, which may affect economic performance.

Durlauf et al. (2012) re-examine BM's study with the goal being to explore replicability and robustness of the finding that religion matters for macroeconomic outcomes. Durlauf et al. (hereafter DKT) report that BM's empirical results are reasonably replicable, but are not robust to model uncertainty with regard to alternative possible growth determinants, a matter that has preoccupied this literature. Extending the framework to a broader set of potential growth models, and allowing for uncertainty with regard to a preferred growth model, DKT find little support for BM's empirical outcome that religion matters for growth. DKT adopt a hybrid frequentist-Bayesian model averaging approach, which allows for over-identification as well as uncertainty with respect to the choice of instrument set and uncertainty with regard to relevant growth factors. They also consider a so-called "kitchen sink" model, which is the largest model in their model space, in the sense of including every proposed explanatory variable.

For our illustration of allowing for possible endogeneity of some of the determinants of growth, we employ one of BM's growth models, along with their instrument choices, using the data provided by DKT[19]; we refer the reader to BM for discussions on the suitability of instrument choices. Specifically, we examine the following model:

$$g_{it} = \sum_{j=1}^{3} \beta_{1j} D_{jt} + \beta_{14} school_{it} + \beta_{15} life_{it} + \beta_{16} fertl_{it} + \beta_{17} totopen_{it} \tag{21}$$

$$+\beta_{18} rule_{it} + \beta_{21} heaven_{it} + \beta_{22} hell_{it} + \beta_{23} chmo_{it} + \beta_{24} easrel_{it} + \beta_{25} hindu_{it}$$

$$+\beta_{26} jews_{it} + \beta_{27} muslim_{it} + \beta_{28} ortho_{it} + \beta_{29} prota_{it} + \beta_{2,10} othrel_{it} + \beta_{2,11} y0_{it}$$

$$+\beta_{2,12} inv_{it} + \beta_{2,13} opres_{it} + \beta_{2,14} gv_{it} + \beta_{2,15} dp_{it} + \beta_{2,16} prights_{it} + \beta_{2,17} prights_{it}^2 + u_{it}.$$

Appendix 1 provides variable definitions, along with adopted instruments. Our specific interest lies in the impacts of the religiosity variables, belief in heaven ($heavenlg$), belief in hell ($helllg$) and monthly church attendance ($chmolg$), as represented by the coefficients β_{21}, β_{22} and β_{23}, respectively. The model also includes regressors that account for adherence to particular religious groups; those linked with the Catholic faith omitted so that $\beta_{24} \dots \beta_{2,10}$ need to be interpreted relative to the Catholic share[20]. The "i" subscript denotes the country, whereas the "t" subscript refers to the period; $t = 1,2,3$, with $t = 1$ corresponding to 1965-1975 and so on. Data issues severely limited the number of possible countries, with BM ultimately including only 41 countries. The dataset provided by DKT only permitted us to examine model (21) for 40 countries (listed in Appendix 1) with some additional exclusions: Hungary, Poland, Bangladesh and Iceland were dropped for the first period; Hungary and Poland were excluded from the second period; and South Africa from the third period. This resulted in $n_1 = 36, n_2 = 38, n_3 = 39$ & $n = 113$. BM estimate model (21) using three stage least squares, whereas, given the nature of our theoretical work, we focus on 2SLS compared with pooled OLS. We recognize the consequences of this on the error term assumptions and subsequent estimates of coefficients and standard errors[21]. Despite these shortcomings, we believe this case provides readers with another illustration of a branch of research for which our theoretical results are likely useful.

[19] DKT provide their data via the Journal of Applied Econometrics Data Archive; see, http://qed.econ.queensu.ca/jae/2012-v27.7/, last accessed 23 March 2017.

[20] BM use the 1980 share for the last period due to the available survey information, justifying this approach due to the persistence in religious faith.

[21] It would be useful to extend our theoretical results to allow for nonspherical errors.

Table 3: Descriptive statistics for some key variables used in Example 2 (pooled sample)

Variable	Minimum	Maximum	Mean	Std. Dev.
	Economic Growth			
Average growth rates of real per capita GDP (%)	-2.1	7.9	2.3	2.0
	Religiosity			
Belief in heaven (%)	19.2	97.0	61.0	20.3
Belief in hell (%)	7.7	95.6	39.3	20.4
Monthly church attendance (%)	9.4	91.3	41.9	22.8
	Religion Shares			
Eastern (%)	0.0	97.0	6.2	22.1
Hindu (%)	0.0	82.7	2.5	13.3
Jewish (%)	0.0	89.6	2.8	14.4
Muslim (%)	0.0	99.3	6.0	19.5
Orthodox (%)	0.0	78.0	2.6	12.5
Protestant (%)	0.1	99.6	26.5	35.5
Other (%)	0.0	46.9	3.4	9.2
Catholic (%)	0.1	99.9	50.0	41.9
	Neoclassical Growth			
Initial income (logged)	6.6	9.7	8.5	0.8
Schooling (years)	0.2	6.0	2.1	1.3
	Demography			
Life expectancy at age 1 (years)	50.8	76.9	69.7	6.3
Fertility rate	1.6	7.4	3.6	1.7

Proceeding, we have $K_1 = 8, K_2 = 17, K = 25, K^* = 42, n = 113, L_2 = 17$; BM classify a substantial number of the explanatory variables as endogenous – we list these in Appendix 1, along with the adopted instruments (which follow BM). Table 3 reports some summary statistics for the relevant explanatory factors. The countries in the sample vary widely in religious beliefs, church attendance and proportions that declare affinity with a particular religious group. Interestingly, the mean country percentage of the population who believe in heaven is over 1.5 times the corresponding statistic for believing in hell. For the countries in our sample, Catholicism and Protestantism are the dominant religions, but our sample also consists of countries whose foremost religion is of other faiths and observations with effectively zero percentage in every religion – identification with one or more of the specified religions differs extensively across the sample.

We present OLS and 2SLS estimates, along with standard errors, in columns (2) and (3) of Table 4, with the other columns providing the estimates from using our weighted WALS estimator for the four considered priors. As in Example 1, an asterisk denotes statistical significance at (at least) the nominal 10% significance level; recall our earlier comment that the assumption of normality for the WALS estimators (even asymptotically) may be unreasonable (even approximately). We also report outcomes from joint tests of non-significance of the religiosity regressors and of the religion shares, providing asymptotic χ^2 p-values, albeit these too may not be accurate. A bold font indicates instances for which the absolute value of the t-ratio is greater than one. Prior to discussing the outcomes reported in the table, we note that the sample DWH statistic value, using the 2SLS estimate of the error variance, is 1.091, with an $F(17,71)$ p-value of 0.380, suggesting that the suspected regressors are not correlated with the error term. For this sample, the pretest estimates would be the OLS estimates.

In discussing the results, we concentrate attention on the religion variables, first the religiosity factors (belief in heaven, belief in hell and monthly church attendance) and secondly the religion share regressors. Comparing the OLS and 2SLS outcomes on the religiosity variables (see columns (2) and (3)), we note substantial variations in magnitudes of coefficient estimates, signs of estimates, standard errors and statistical significance. For instance, the estimated 2SLS impact of monthly church attendance on average growth rate is more than twice that obtained using OLS, and OLS suggests that belief in hell (belief in heaven) is (not) statistically significant, whereas 2SLS detects the reverse outcome. Standard errors using 2SLS are more than five times higher than those from OLS. Perhaps not surprisingly given these disparities, the Wald test of joint non-significance of the relevance of the religiosity variables is rejected with OLS but not so with 2SLS. The comparable outcomes from BM (using their sample and 3SLS) is column (6) in their Table 4 (p773) and from DKT (using a similar sample to ours) we refer the reader to column (6) of their Table II (p1064). That we find some dissimilarities between our results from those reported by BM/DKT highlights the sensitivity of outcomes to the selected sample, estimation method, and that we do not report robust standard errors - the conclusions of BM are often not robust to such matters.

The $\hat{\lambda}$ estimates are similar across the four examined priors, only varying by at most 8% with the Laplace $\hat{\lambda}$ being the highest estimate and the Subbotin ($q = 0.5$) the lowest; the weights for the Subbotin ($q = 0.7995$) and the Weibull methods are the same (at least to the reported number of decimal places). The 2SLS estimates are weighted more than the OLS ones across the four WALS cases, and the resulting coefficient estimates for the religiosity factors are practically the same. The standard errors obtained using WALS are significantly higher than those from OLS, but at least 16% lower than those from 2SLS. As the standard errors for the Subbotin ($q = 0.5$) prior method are slightly smaller than for the other priors, we observe that this is the one situation for which the three religiosity variables have a t-ratio greater than one in magnitude – this outcome occurs only for belief in heaven and monthly church attendance with the other priors. Interpreting the scenario of an absolute t-ratio greater than one as indicating that a regressor is sufficiently correlated with average growth rates, these WALS estimates imply that economic growth responds positively to belief in heaven and negatively to monthly church attendance, qualitative conclusions that accord with those generally arrived at by BM. Of the religiosity regressors, monthly church attendance seems to be the most important variable, followed by belief in heaven.

While our results regarding monthly church attendance support the outcomes of BM, this is not the case regarding belief in hell. BM report that economic growth and belief in hell is significantly positively correlated, whereas we only detect this relationship using OLS, with the WALS results hinting at a negative association, if there is any relevant connection. DKT, on the other hand, detect little evidence of religiosity mattering for growth in their robustness study that changes the fundamental model specification and employs Bayesian model averaging to incorporate the variety of model specifications. If religion demonstrates any effect on economic growth, DKT's work alludes to monthly church attendance possibly leading to lower economic growth rather than higher growth as suggested by BM. It remains for future work to ascertain whether our WALS method would lead to comparable conclusions should we extend our approach in the directions examined by DKT.

Turning to the effects of the religion adherence shares, which should be interpreted as the impact relative to that for Catholicism, BM show (with their 3SLS estimator and sample) that the statistically significant, negative, factors are for Hindu, Muslim, Orthodox and Protestant. Our OLS results only support this empirical conclusion for Muslim and Protestant, with Other Religions also indicating a significant, negative effect. In contrast, the 2SLS outcomes detect that only Hindu is statistically significant with a positive impact (relative to Catholicism). We see that the numerical estimates for these faith factors can markedly vary across OLS and 2SLS, along with standard errors that are much higher for 2SLS. The hybrid approach of averaging 2SLS and OLS using WALS results in similar numerical estimates and standard errors for the religion shares impacts on growth across the four examined priors, with the Subbotin ($q = 0.5$) standard errors being marginally lower if at all. The WALS standard errors are noticeably lower than those from 2SLS, but still considerably higher than from OLS. Only Protestant and Other Religions have t-ratios larger than one in magnitude, the rough standard we are adopting with WALS, with Protestant resulting in the highest absolute t-ratio. Given these outcomes, that the joint effect of the religion affiliation shares on economic growth is non-significant is as expected, although we note again that the reasonableness of the p-values reported for the WALS cases is unknown. That the religion shares are jointly non-significant conflicts with BM's conclusions but accord with those qualitatively reported by DKT, at least for their model averaging results.

As with Example 1, this second example illustrates that the WALS model-averaging approach provides a strategy that readily produces estimates of the coefficients and standard errors that are not conditional on an assumption regarding endogeneity of a subset of regressors, taking account of any such uncertainty.

Table 4: Estimated average growth rates equation

Explanatory Variable	OLS	2SLS	WALS			
			Laplace $q = 1$ $c = log2$	Subbotin $q = 0.5$ $c = 1.6783$	Subbotin $q = 0.7995$ $c = 0.9377$	Weibull $q = 0.8876$ $c = log2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
D_1	$\mathbf{0.338}^*$ $\mathbf{(0.068)}$	$\mathbf{0.525}^*$ $\mathbf{(0.186)}$	$\mathbf{0.456}^*$ $\mathbf{(0.166)}$	$\mathbf{0.447}^*$ $\mathbf{(0.163)}$	$\mathbf{0.451}^*$ $\mathbf{(0.166)}$	$\mathbf{0.451}^*$ $\mathbf{(0.166)}$
D_2	$\mathbf{0.327}^*$ $\mathbf{(0.068)}$	$\mathbf{0.515}^*$ $\mathbf{(0.187)}$	$\mathbf{0.446}^*$ $\mathbf{(0.167)}$	$\mathbf{0.437}^*$ $\mathbf{(0.164)}$	$\mathbf{0.441}^*$ $\mathbf{(0.166)}$	$\mathbf{0.441}^*$ $\mathbf{(0.166)}$
D_3	$\mathbf{0.327}^*$ $\mathbf{(0.067)}$	$\mathbf{0.511}^*$ $\mathbf{(0.183)}$	$\mathbf{0.443}^*$ $\mathbf{(0.163)}$	$\mathbf{0.434}^*$ $\mathbf{(0.160)}$	$\mathbf{0.439}^*$ $\mathbf{(0.163)}$	$\mathbf{0.439}^*$ $\mathbf{(0.163)}$
school	$\mathbf{0.0044}^*$ $\mathbf{(0.0019)}$	$\mathbf{0.0059}^*$ $\mathbf{(0.0036)}$	$\mathbf{0.0053}^*$ $\mathbf{(0.0034)}$	$\mathbf{0.0053}^*$ $\mathbf{(0.0034)}$	$\mathbf{0.0053}^*$ $\mathbf{(0.0034)}$	$\mathbf{0.0053}^*$ $\mathbf{(0.0034)}$
life	$\mathbf{-0.051}^*$ $\mathbf{(0.027)}$	-0.029 (0.059)	-0.037 (0.054)	-0.038 (0.052)	-0.038 (0.054)	-0.038 (0.054)
fertl	$\mathbf{-0.011}^*$ $\mathbf{(0.008)}$	$\mathbf{-0.050}^*$ $\mathbf{(0.027)}$	$\mathbf{-0.035}^*$ $\mathbf{(0.024)}$	$\mathbf{-0.034}^*$ $\mathbf{(0.024)}$	$\mathbf{-0.035}^*$ $\mathbf{(0.024)}$	$\mathbf{-0.035}^*$ $\mathbf{(0.024)}$
totopen	$\mathbf{0.118}$ $\mathbf{(0.102)}$	0.056 (0.196)	0.079 (0.186)	0.082 (0.182)	0.080 (0.186)	0.080 (0.185)
rule	$\mathbf{0.014}^*$ $\mathbf{(0.011)}$	0.006 (0.026)	0.009 (0.024)	0.009 (0.023)	0.009 (0.023)	0.009 (0.023)
heaven	0.0032 (0.0046)	$\mathbf{0.056}^*$ $\mathbf{(0.037)}$	$\mathbf{0.036}$ $\mathbf{(0.031)}$	$\mathbf{0.034}^*$ $\mathbf{(0.030)}$	$\mathbf{0.035}$ $\mathbf{(0.031)}$	$\mathbf{0.035}$ $\mathbf{(0.031)}$
hell	$\mathbf{0.0074}^*$ $\mathbf{(0.0051)}$	-0.025 (0.031)	-0.013 (0.025)	$\mathbf{-0.012}$ $\mathbf{(0.021)}$	-0.012 (0.025)	-0.012 (0.025)
chmo	$\mathbf{-0.011}^*$ $\mathbf{(0.004)}$	$\mathbf{-0.028}^*$ $\mathbf{(0.020)}$	$\mathbf{-0.022}^*$ $\mathbf{(0.016)}$	$\mathbf{-0.021}^*$ $\mathbf{(0.016)}$	$\mathbf{-0.021}^*$ $\mathbf{(0.016)}$	$\mathbf{-0.021}^*$ $\mathbf{(0.016)}$
easrel	0.0037 (0.0096)	0.0073 (0.057)	0.0059 (0.026)	0.0058 (0.023)	0.0059 (0.025)	0.0059 (0.025)
hindu	-0.0090 (0.0156)	$\mathbf{0.076}^*$ $\mathbf{(0.057)}$	0.045 (0.050)	0.041 (0.050)	0.043 (0.050)	0.043 (0.050)
jews	0.0067 (0.0127)	$\mathbf{0.032}$ $\mathbf{(0.031)}$	0.023 (0.028)	0.021 (0.027)	0.022 (0.028)	0.022 (0.028)
muslim	$\mathbf{-0.015}$ $\mathbf{(0.012)}$	-0.0083 (0.043)	-0.011 (0.036)	-0.011 (0.029)	-0.011 (0.036)	-0.011 (0.036)
ortho	-0.144 (0.015)	-0.037 (0.050)	-0.029 (0.042)	-0.028 (0.039)	-0.028 (0.042)	-0.028 (0.042)
prota	$\mathbf{-0.012}^*$ $\mathbf{(0.007)}$	$\mathbf{-0.032}$ $\mathbf{(0.028)}$	$\mathbf{-0.025}$ $\mathbf{(0.023)}$	$\mathbf{-0.024}$ $\mathbf{(0.022)}$	$\mathbf{-0.024}$ $\mathbf{(0.023)}$	$\mathbf{-0.024}$ $\mathbf{(0.023)}$
othrel	$\mathbf{-0.023}$ $\mathbf{(0.021)}$	$\mathbf{-0.076}$ $\mathbf{(0.066)}$	$\mathbf{-0.056}$ $\mathbf{(0.056)}$	$\mathbf{-0.054}$ $\mathbf{(0.054)}$	$\mathbf{-0.055}$ $\mathbf{(0.055)}$	$\mathbf{-0.055}$ $\mathbf{(0.055)}$
y0	$\mathbf{-0.029}^*$ $\mathbf{(0.005)}$	$\mathbf{-0.049}^*$ $\mathbf{(0.017)}$	$\mathbf{-0.042}^*$ $\mathbf{(0.015)}$	$\mathbf{-0.041}^*$ $\mathbf{(0.015)}$	$\mathbf{-0.041}^*$ $\mathbf{(0.015)}$	$\mathbf{-0.041}^*$ $\mathbf{(0.015)}$
inv	0.021 (0.035)	0.014 (0.087)	0.017 (0.077)	0.017 (0.075)	0.017 (0.076)	0.017 (0.076)
opres	$\mathbf{0.033}^*$ $\mathbf{(0.009)}$	$\mathbf{0.028}$ $\mathbf{(0.024)}$	$\mathbf{0.030}^*$ $\mathbf{(0.021)}$	$\mathbf{0.030}^*$ $\mathbf{(0.020)}$	$\mathbf{0.030}^*$ $\mathbf{(0.021)}$	$\mathbf{0.030}^*$ $\mathbf{(0.021)}$
gv	0.021 (0.046)	-0.176 (0.187)	-0.104 (0.159)	-0.095 (0.151)	-0.100 (0.158)	-0.100 (0.158)
dp	$\mathbf{-0.011}^*$ $\mathbf{(0.005)}$	-0.033 (0.034)	-0.025 (0.028)	$\mathbf{-0.024}$ $\mathbf{(0.026)}$	-0.025 (0.027)	-0.025^* (0.027)
prights	0.020 (0.037)	-0.054 (0.134)	-0.026 (0.111)	-0.023 (0.105)	-0.025 (0.109)	-0.025 (0.109)
prights^2	-0.015 (0.030)	0.037 (0.101)	0.018 (0.084)	0.015 (0.080)	0.017 (0.083)	0.017 (0.083)
$\hat{\lambda}$	*n.a.*	*n.a.*	0.630	0.585	0.608	0.608
H_0^1	10.223^* (0.017)	3.332 (0.343)	2.211 (0.530)	2.209 (0.530)	2.124 (0.547)	2.111 (0.550)
H_0^2	9.387 (0.226)	4.575 (0.712)	3.634 (0.821)	3.509 (0.834)	3.547 (0.830)	3.532 (0.832)

Notes: Standard errors are provided in parentheses. A superscript * indicates statistical significance, against an appropriate one-sided alternative hypothesis (as relevant) with, at least, an approximate 10% significance level. The null hypotheses reported in the table relate to: $H_0^1 = \beta_{21} = \beta_{22} = \beta_{23} = 0$ and $H_0^2 = \beta_{24} = \beta_{25} = \cdots = \beta_{2,10} = 0$, where the presented numbers are sample Wald statistics with, respectively, $\chi^2(3)$ and $\chi^2(7)$ p-values in parentheses. A bold font indicates a t-ratio greater than one in magnitude.

6. Concluding Remarks

This paper has three purposes. First, it introduces a model averaging estimator, within the framework of weighted-average least squares, which combines the OLS and 2SLS estimators when there is ambiguity about whether one or more regressors are endogenous. Our approach allows for a proper treatment of such model uncertainty, rather than the common frequentist method of reporting either the OLS or 2SLS results (or both), or undertaking a preliminary test to ascertain which estimator might be preferred, moving the researcher into a pretest framework, which is more often than not ignored when conveying the empirical outcomes. Accounting for the model selection is essential if we are genuinely concerned with reporting appropriate results. Our consideration of the issue focuses on obtaining the best possible estimates of the model parameters, allowing for the model uncertainty, rather than choosing between either OLS or 2SLS, with the WALS structure providing for both the model that assumes no endogenous regressors and the model that allows endogeneity to provide knowledge. Via a number of theorems and corollaries, we extend the current WALS theory to develop a mixed OLS/2SLS estimator that combines model selection and estimation, deriving the asymptotic bias, variance and risk functions of the estimator.

Our second objective is to show that the qualitative findings from the existing WALS literature equally applies to the asymptotic properties of our considered estimator. Specifically, ascertaining how best to estimate the mean of a normally distributed random vector, whose components are independently distributed, crucially determines the asymptotic properties of our mixed OLS/2SLS estimator. That is, we only need establish a preferred way to estimate the mean of a standard normal random variable. Having resolved this matter, according to some chosen criteria, the subsequently obtained weight is also optimal for averaging the OLS and 2SLS estimators of the coefficient vector of our structural model.

Our third intention is to illustrate our model averaging WALS estimator, utilizing previous WALS research. We adopt a Bayesian approach to form the required estimator of the mean of a normally distributed random variate, so that the adopted estimator is a mixed Bayesian-frequentist rule. The priors we consider are “neutral”, in the sense of attempting to simulate ignorance about model preference, and have been shown elsewhere as having some other favourable properties. We consider two empirical examples, re-examining the results reported in the original work. Our first example, taken from a popular undergraduate econometrics textbook (Wooldridge, 2016) based on a model and data from Card (1995), investigates the link between education and earnings. Our second demonstration re-examines a model considered by Barro and McCleary (2003), which explores the impact of religious beliefs (belief in heaven and belief in hell), involvement in religious activities (proxied by monthly church attendance) and observance of particular religions. We consider four priors considered by the WALS research to date. Both examples illustrate that WALS can be readily applied in practice, and is a viable model averaging estimator when model uncertainty exists regarding the exogeneity assumption concerning one or more explanatory variables.

In terms of future work, simulation studies would be helpful to explore the finite-sample properties of the OLS-2SLS WALS model averaging estimator, especially compared with the usually adopted pretest estimator. Another topic would be the question of prior choice. We investigate priors proposed by WALS works focusing on the issue of estimating the parameters for a set of "focus" explanatory variables, deemed necessary for inclusion in the model, when there is another group of auxiliary factors that may or may not be relevant, whose parameters are not of fundamental interest. In such an environment, using a prior that is "neutral" or "ignorant" with regard to whether or not to include one or more of the auxiliary variables (more specifically, is ignorant to the issue of whether the associated t-ratio is larger or smaller than one in absolute value) makes eminent sense. However, as the researcher may wish to adopt an informative prior regarding the endogeneity question, it would be interesting to examine such a matter.

Additionally, our setup assumes that the specification of the explanatory variables is reasonable, whereas uncertainty about this issue is more the norm. Moreover, theoretical doubt may also exist regarding instrument choices, an assumption that we do not query. Systematically extending our research to allow for both of these concerns would be worthy of attention.

References

Barro, R. and R. McCleary, 2003, Religion and economic growth across countries, American Sociological Review 68, 760-781.

Barro, R. and X. Sala-i-Martin, 2004, Economic Growth 2nd Edition (MIT Press, Cambridge, MA).

Baum, C.F., M.E. Schaffer and S. Stillman, 2003, Instrumental variables and GMM: Estimation and testing, The Stata Journal 3, 1-31.

Card, D., 1995, Using geographic variation in college proximity to estimate the returns to schooling, in Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp. Eds. Christophides, L.N., E.K. Grant and R. Swidinsky (University of Toronto Press, Toronto), 201-222.

Claeskens, G. and N.L. Hjort, 2008, Model Selection and Model Averaging (Cambridge University Press, Cambridge, U.K.).

Clarke, J.A., 2008, On weighted estimation in linear regression in the presence of parameter uncertainty, Economics Letters 100, 1-3.

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 J.R. Magnus, 2004, On the harm that ignoring pretesting can cause, Journal of Econometrics 122, 27-46.

De Luca, G. and J.R. Magnus, 2011, Bayesian model averaging and weighted-average least squares: Equivalence, stability, and numerical issues, The Stata Journal 11, 518-544.

De Lucca, G., J.R. Magnus and F. Perrachi, 2017, Weighted-average least squares estimation of generalized linear papers, Working Paper, available at http://www.janmagnus.nl/wips/wip-dlmp1.pdf, last accessed 8 March 2017.

Durlauf, S.N., A. Kourtellos and C.M. Tan, 2008, Are any growth theories robust?, The Economic Journal 18, 329-346.

Durlauf, S.N., A. Kourtellos and C.M. Tan, 2012, Is God in the details? A re-examination of the role of religion in economic growth, Journal of Applied Econometrics 27, 1059-1075.

Durbin, J., 1954, Errors in variables, Review of the International Statistical Institute 22, 23-32.

Einmahl, J., J.R. Magnus and K. Kumar, 2011, On the choice of prior in Bayesian model averaging, CentER Discussion Paper, Vol. 2011-003, Tilburg University: Econometrics.

Giles, J.A. and D.E.A. Giles, 1993, Pre-test estimation and testing in econometrics: Recent developments, Journal of Economic Surveys 7, 145-197.

Gourieroux, C. and A. Trognon, 1984, Specification pre-test estimator, Journal of Econometrics 25, 15-27.

Hansen, B.E., 2007, Least squares model averaging, Econometrica 75, 1175-1189.

Hausman, J.A., 1978, Specification tests in econometrics, Econometrica 46, 1251-1271.

Hjort, N.L. and G. Claeskens, 2003a, Frequentist model average estimators (with discussion), Journal of the American Statistical Association 98, 879-899.

Hjort, N.L. and G. Claeskens, 2003b, Rejoinder to "Frequentist model average estimators" and "The focused information criterion", Journal of the American Statistical Association 98, 938-945.

Hoeting, J.A., Madigan, D., Raftery, A.E. and C.T. Volinsky, 1999, Bayesian model averaging: A tutorial (with discussion), Statistical Science 14, 382-417.

James, W. and C. Stein, 1960, Estimation with quadratic loss, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1 (University of California Press, Berkeley), 197-206.

Judge, G.G. and M.E. Bock, 1978, The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics (North-Holland, Amsterdam).

Kim, T.H. and H. White, 2001, James-Stein-type estimators in large samples with application to the least absolute deviations estimator, Journal of the American Statistical Association 96, 697-705.

Kumar, K. and J.R. Magnus, 2013, A characterization of Bayesian robustness for a normal location parameter, Sankhya (Series B) 75, 216-237.

Liang, H., G. Zou, A.T.K. Wan and X. Zhang, 2011, Optimal weight choice for frequentist model average estimators, Journal of the American Statistical Association 106, 1053-1066.

Lenkoski, A., T.S. Eicher and A.E. Raftery, 2014, Two-stage Bayesian model averaging variable models, Econometric Reviews 33, 122-151.

Magnus, J.R., 1999, The traditional pretest estimator, Theory of Probability and Its Applications 55, 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 G. De Luca, 2016, Weighted-average least squares (WALS): A survey, Journal of Economic Surveys 30, 117-148.

Magnus, J.R. and J. Durbin, 1999, Estimation of regression coefficients of interest when other regression coefficients are of no interest, Econometrica 67, 639-643.

Magnus, J.R., O. Powell and P. Prüfer, 2010, A comparison of two model averaging techniques with an application to growth empirics, Journal of Econometrics 154, 139-153.

Magnus, J.R., A.T.K. Wan and X. Zhang, 2011, Weighted-average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market, Computational Statistics and Data Analysis 55, 1331-1341.

Magnus, J.R. and W. Wang, 2014, Concept-based Bayesian model averaging and growth empirics, Oxford Bulletin of Economics and Statistics 76, 874-897.

Mander, A., 2012, INTEGRATE: Stata module to perform one-dimensional integration, Statistical Software Components S457429, Boston College: Economics.

Moral-Benito, E., 2015, Model averaging in economics: An overview, Journal of Economic Surveys 29, 46-75.

Morey, M.J., 1984, The statistical implications of preliminary specification error testing, Journal of Econometrics 25, 63-72.

Nakamura, A. and M. Nakamura, 1981, On the relationships among several specification error tests presented by Durbin, Wu and Hausman, Econometrica 49, 1583-1588.

Poghosyan, K. and J.R. Magnus, 2012, WALS estimation and forecasting in factor-based dynamic models with an application to Armenia, International Econometric Review 4, 40-58.

Raftery, A.E. and Y. Zheng, 2003, Discussion to “Frequentist model average estimators” and “The focused information criterion” by Hjort, N.L. and Claeskins, G., Journal of the American Statistical Association 98, 931-938.

Raftery, A.E., D. Madigan and J.A. Hoeting, 1997, Bayesian model averaging for linear regression models, Journal of the American Statistical Association 92, 179-191.

Richardson, D.H. and D.M. Wu, 1971, A note on the comparison of ordinary and two-stage least squares estimators, Econometrica 39, 973-981.

Sala-i-Martin, X., G. Doppelhofer and R.I. Miller, 2004, Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach, American Economic Review 94, 813-835.

Skeels, C.L. and L.W. Taylor, 1995, On a simultaneous equations pre-test estimator, Journal of Econometrics 68, 269-286.

StataCorp., 2015, Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

Toro-Vizcarrondo, C. and T.D. Wallace, 1968, A test of the mean square error criterion for restrictions in linear regression, Journal of the American Statistical Association 63, 558-572.

Wooldridge, J.M., 2016, Introductory Econometrics: A Modern Approach 6th Edition (Cengage Learning, Boston, MA).

Wu, D.M., 1973, Alternative tests of independence between stochastic regressors and disturbances, Econometrica 41, 733-750.

Zou, G., A.T.K. Wan, X. Wu and T. Chen, 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.

Appendix 1: Information for the Applications

Example 1: Returns to Schooling

We specify the sources of the data for this illustration in Section 5.1 and footnote 21. Definitions of the variables used in model (20) are:

lwage	= natural logarithm of *wage*.
wage	= hourly wage in cents in 1976.
educ	= years of schooling in 1976.
age	= age in years in 1976.
exper	= *age* – *educ* – 6.
expersq	= $(exper)^2$.
black	= 1 if classified as Black, 0 otherwise.
smsa	= 1 if resided in Standard Metropolitan Statistical Area (SMSA) in 1976, 0 otherwise.
smsa66	= 1 if resided in SMSA in 1966, 0 otherwise.
south	= 1 if lived in a classified Southern region in 1976, 0 otherwise.
nearc4	= 1 if lived near a 4 year college in 1966, 0 otherwise.
reg66j	= 1 if resided in region j in 1966, 0 otherwise.

Example 2: Role of Religion in Economic Growth

We obtained the data for this case as detailed in Section 5.1 and footnote 18. BM and DKT provide specific sources. The variables, and instruments, used to estimate model (21) are as follows:

g	= average growth rates for the periods 1965-75, 1975-85, 1985-95.
D_j	= period dummy variables for 1965-75, 1975-85, 1985-95.
y0	= natural logarithm of per capita GDP at 1965, 1975, 1985. The instrument is the value of *y0* at 1960, 1970, 1980.
gpop	= natural logarithm of average population growth rates plus 0.05 for the periods 1965-75, 1975-85, 1985-95. The instrument is the average values of 1960-65, 1970-75, 1980-85.
inv	= natural logarithm of average ratios over each period of investment for the periods 1965-75, 1975-85, 1985-95. The instrument is the average values of 1960-65, 1970-75, 1980-85.
school	= years of male secondary and higher school attainment in 1965, 1975, 1985.
life	= reciprocals of life expectancy at age 1 in 1960, 1970, 1980.
fertl	= natural logarithm of the total fertility rate in 1960, 1970, 1980.
opres	= average ratios for each period of exports plus imports to GDP, filtered for the usual relation of this ratio to the natural logarithm of population and area for the periods 1965-75, 1975-85, 1985-95. The instrument is the average values of 1960-65, 1970-75, 1980-85.
gv	= average ratios of government consumption (net of outlays on defense and education) to GDP for the periods 1965-75, 1975-85, 1985-95. The instrument is the average values of 1960-65, 1970-75, 1980-85.
dp	= the consumer price inflation rate for the periods 1965-75, 1975-85, 1985-95. The instrument is *spainpor*.
totopen	= growth rate of the terms of trade over each period, interacted with the average ratio of exports plus imports to GDP.
hell	= fraction of the population who believe in hell expressed in the form of $\log\left(\frac{x}{1-x}\right)$. See the note below for instrument list.
heaven	= fraction of the population who believe in heaven expressed in the form of $\log\left(\frac{x}{1-x}\right)$. See the note below for instrument list.

chmo = population averages of monthly church attendance expressed in the form of $\log\left(\frac{x}{1-x}\right)$. See the note below for instrument list.

easrel = Eastern religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. This category includes Buddhism, Chinese Universists, Confucians, Neoreligionists, Shintos, and Zoroastrians. The instrument is *easrel* in 1970, 1970, 1980.

hindu = Hindu religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. The instrument is *hindu* in 1970, 1970, 1980.

jew = Jewish religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. The instrument is *jew* in 1970, 1970, 1980.

muslim = Muslim religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. The instrument is *muslim* in 1970, 1970, 1980.

ortho = Orthodox religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. The instrument is *ortho* in 1970, 1970, 1980.

prota = Protestant religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. The instrument is *prota* in 1970, 1970, 1980.

othrel = Other Religion share in 1970, 1980, 1980, expressed as a fraction of the population who expressed this religious faith. The instrument is *othrel* in 1970, 1970, 1980.

prights = political rights. The average for each period of the Freedom House measure of democracy (electoral rights). The average of 1972-74 is used for the first period. The instrument is lagged *prights*.

relplu = religious pluralism, defined as one minus the Herfindahl Index, with this index defined as the estimated probability that two randomly selected persons from the population belong to the same religion, in 1970, 1980, 1980 (1990 for some countries). See BM's footnote 11 (p764) for further details.

state_rel70 = 1 if there existed a state religion in 1970, 0 otherwise.

state_reg70 = 1 if there existed state regulation of religion in 1970, 0 otherwise.

spainpor = 1 if the country was colonized by Spain or Portugal, 0 otherwise.

Note: The explanatory variables *hell, heaven* and *chmo* are jointly instrumented by the variables *state_rel70*, *state_reg70* and *relplu*.

List of countries for Example 2:

North America: Canada, United States

Europe: Austria, Belgium, Cyprus, Switzerland, Denmark, Spain, Finland, France, United Kingdom, Hungary, Ireland, Iceland, Israel, Italy, Netherlands, Norway, Poland, Portugal, Sweden, Turkey

Asia and Oceania: Australia, India, Bangladesh, Japan, Republic of Korea, New Zealand, Philippines, Taiwan

Sub-Saharan Africa: Ghana, South Africa

Latin America & Caribbean: Argentina, Brazil, Chile, Dominican Republic, Mexico, Peru, Uruguay, Venezuela