THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 15-38

Claudio Morana
University of Milan-Bicocca, Italy
The Rimini Centre for Economic Analysis, Italy

MODEL AVERAGING BY STACKING

The *Rimini Centre for Economic Analysis* (RCEA) was established in March 2007. RCEA is a private, nonprofit organization dedicated to independent research in Applied and Theoretical Economics and related fields. RCEA organizes seminars and workshops, sponsors a general interest journal *The Review of Economic Analysis*, and organizes a biennial conference: *The Rimini Conference in Economics and Finance* (RCEF) . The RCEA has a Canadian branch: *The Rimini Centre for Economic Analysis in Canada* (RCEA-Canada). Scientific work contributed by the RCEA Scholars is published in the RCEA Working Papers and Professional Report series.

The Rimini Centre for Economic Analysis
Legal address: Via Angherà, 22 – Head office: Via Patara, 3 - 47921 Rimini (RN) – Italy
www.rcfea.org - secretary@rcfea.org

Model Averaging by Stacking

Claudio Morana

Department of Economics, Management and Statistics, University of Milan-Bicocca, Milan, ITALY
Email: claudio.morana@unimib.it

Center for Research on Pensions and Welfare Policies, Collegio Carlo Alberto, Moncalieri, ITALY
Rimini Center for Economic Analysis, Rimini, ITALY

October 2015

Abstract

The paper introduces a new Frequentist model averaging estimation procedure, based on a stacked OLS estimator across models, implementable on cross-sectional, panel, as well as time series data. The proposed estimator shows the same optimal properties of the OLS estimator under the usual set of assumptions concerning the population regression model. Relatively to available alternative approaches, it has the advantage of performing model averaging ex-ante in a single step, optimally selecting models' weight according to the MSE metric, i.e., by minimizing the squared Euclidean distance between actual and predicted value vectors. Moreover, it is straightforward to implement, only requiring the estimation of a single OLS augmented regression. By exploiting ex-ante a broader information set and benefiting of more degrees of freedom, the proposed approach yields more accurate and (relatively) more efficient estimation than available ex-post methods.

Keywords

Model Averaging; Model Uncertainty

1. Introduction

The Classical Linear Regression Model (CLRM) is grounded on a basic set of assumptions concerning its specification and distributional properties of control variables and error term. In this respect, under what is usually held as Assumption 1, the population regression model is required to be linear in the parameters, and control variables are known and all included in the model. However, the latter *correct specification* assumption might not always be appropriate in Economics; for instance, there might be more than a single set of variables, i.e., more than a single candidate model, which could be employed in estimation, also when economic theory has clear-cut implications for the causal linkage of interest.

Think of the relationship linking y to x, when both variables can be measured in different ways, i.e., when there exist y_i and x_j, $i = 1, ..., P$, $j = 1, ..., R$; then, in principle, up to $P \times R$ different models could

be estimated.[1]

Two solutions have so far been proposed in the literature to the above model selection problem. On the one hand, by maintaining the assumption of correct specification, a selection of a single model out of the $P \times R$ candidates can be performed, based on various specification strategies (see [2] for a general account; see also [3] for recent developments in model selection). Alternatively, all of the $P \times R$ models can be estimated, and a weighted average across models computed ex-post for the parameters of interest. In the latter case, the assumption of correct specification does not have necessarily to be maintained.

Several model averaging procedures have been proposed in the literature, making use of either Bayesian or Frequentist procedures (see [4], [5]). Admittedly, relatively to Bayesian, the Frequentist approach to model averaging is fairly underdeveloped. The current paper then aims at filling this gap in the literature, by proposing an ex-ante, Mean Square Error-optimal, model averaging procedure. The proposed procedure is grounded on a stacked OLS estimator across models, implementing model averaging ex-ante in a single step, optimally selecting models' weight according to the MSE metric, i.e., by minimizing the squared Euclidean distance between actual and predicted value vectors. Moreover, it is straightforward to compute, only requiring the estimation of a single OLS augmented regression. By exploiting a broader information set ex-ante, i.e., by making use of all the available information jointly, and benefiting of more degrees of freedom, the proposed estimator then yields more accurate and (relatively) more efficient estimation than available ex-post methods. Extension to other estimation frameworks, i.e., GIVE or GMM, is also straightforward.

The rest of the paper is organized as follows. In Section 2 the proposed approach is illustrated by means of a simple example. Then, the econometric methodology is outlined in full in Section 3, while Section 4 deals with its statistical properties. Finally Section 5 concludes.

2. Ex-ante model averaging: An example

For sake of clarity, consider the following bivariate example

$$y_t = \beta x_t + \varepsilon_t \tag{1}$$

where the dependent variable y is a linear function of the independent variable x. The endogenous variable y can then be alternatively measured by y_1 and y_2, while the independent variable x by x_1 and x_2. In what follows we assume that the other usual properties of the CLRM hold, i.e., $\{y_{i,t}, x_{j,t}\}$, $i, j = 1, 2$, $t = 1, ..., T$, $T > 1$, is a stationary and ergodic process, of zero mean for simplicity; the regressors $x_{j,t}$ and the residuals $\varepsilon_{ij,t}$ are at least contemporaneously orthogonal, i.e., $E[\varepsilon_{i,t}|\mathbf{x}_{j,t}] = 0$; the residuals are conditionally homoskedastic ($E\left[\varepsilon_{i,t}^2|x_{t,j}\right] = \sigma^2$) and non serially correlated ($E[\varepsilon_{i,t}\varepsilon_{i,t-n}|x_{t,j}] = 0$, $n = 1, ...$).[2]

Four consistent estimates of the parameter of interest β are then obtained, i.e., $\hat{\beta}_{1,1}$, $\hat{\beta}_{1,2}$, $\hat{\beta}_{2,1}$, $\hat{\beta}_{2,2}$, by means of OLS estimation of each of the four available alternative models

$$\begin{aligned} y_{1,t} &= \beta x_{1,t} + \varepsilon_{11,t} \\ y_{1,t} &= \beta x_{2,t} + \varepsilon_{12,t} \\ y_{2,t} &= \beta x_{1,t} + \varepsilon_{21,t} \\ y_{2,t} &= \beta x_{2,t} + \varepsilon_{22,t}. \end{aligned} \tag{2}$$

Ex-post model averaging then yields a robust consistent estimates $\hat{\beta}_{ep}$ of β, by computing a weighted average of the four available estimates $\hat{\beta}_{1,1}$, $\hat{\beta}_{1,2}$, $\hat{\beta}_{2,1}$, $\hat{\beta}_{2,2}$, with weights determined according to Bayesian or Frequentist approaches.

[1] In Economics the above situation is not unusual. For instance, be y a measure of income distribution inequality and x the degree of financial development of a country; in the latter case, the Gini Index, net or gross, or top-to-bottom income distribution quantile ratios (top to bottom 1% or 10%) would all be valid candidate dependent variables; moreover, concerning financial deepening, the GDP share of liquidity (M2 or M3), stock market capitalization, or credit to the private sector, might be alternatively employed (see [1]).

[2] t is not necessarily a temporal index; applications to cross-sectional data are as viable as to time series.

For instance, within a Frequentist model averaging approach [2], one has

$$\hat{\beta}_{ep} = \sum_{i=1,2} \sum_{j=1,2} \tilde{w}_{i,j} \hat{\beta}_{i,j} \tag{3}$$

where the weights $\tilde{w}_{ij}$ can be computed by means of information criteria as in [6], setting

$$\tilde{w}_{i,j} = \frac{\exp(-I_{i,j}/2)}{\sum_{i=1,2} \sum_{j=1,2} \exp(-I_{i,j}/2)} \tag{4}$$

where $I_{i,j}$ is the Akaike or Schwarz-Bayes information criterion for model i, j. Other approaches are also available, based on Mallow's criterion [7] or cross-validation [8].

On the other hand, the proposed model averaging strategy is single-step and implemented by means of an augmented regression model using all the available data jointly. It then requires the construction of the auxiliary dependent (y_J) and independent (x_J) variables, by appropriately stacking the actual data y_i and x_j in single column vectors.

With reference to the set of models in (2), consider the stacked model obtained from their union, i.e.,

$$\mathbf{y}_J = \mathbf{x}_J \beta + \boldsymbol{\varepsilon}_J \tag{5}$$

where $\mathbf{y}_J = \left[\begin{array}{cccc} \mathbf{y}_1' & \mathbf{y}_1' & \mathbf{y}_2' & \mathbf{y}_2' \end{array} \right]'$, $\mathbf{x}_J = \left[\begin{array}{cccc} \mathbf{x}_1' & \mathbf{x}_2' & \mathbf{x}_1' & \mathbf{x}_2' \end{array} \right]'$ and $\boldsymbol{\varepsilon}_J = \left[\begin{array}{cccc} \boldsymbol{\varepsilon}_{1,1}' & \boldsymbol{\varepsilon}_{1,2}' & \boldsymbol{\varepsilon}_{2,1}' & \boldsymbol{\varepsilon}_{2,2}' \end{array} \right]'$ are $S \times 1$ vectors, $S = 4T$; $\mathbf{y}_i$, $\mathbf{x}_j$, $\boldsymbol{\varepsilon}_{i,j}$, $i, j = 1, 2$, are $T \times 1$ vectors containing the observations on y_i, x_j and $\varepsilon_{i,j}$, respectively.

Alternatively, the regression model can be written as

$$y_{J,s} = \beta x_{J,s} + \varepsilon_{J,s}, \;\; s = 1, ..., S \tag{6}$$

The stacked OLS problem is then stated as

$$\min_{\hat{\beta}_{ea}} RSS(\hat{\beta}_{ea}) \equiv \sum_{s=1}^{S} \left(y_{J,s} - \hat{\beta}_{ea} x_{J,s} \right)^2 \tag{7}$$

yielding, after some algebra,

$$\hat{\beta}_{ea} = \frac{\sum_{i=1,2} \sum_{j=1,2} \sum_{t=1}^{T} y_{i,t} x_{j,t}}{\sum_{i=1,2} \sum_{j=1,2} \sum_{t=1}^{T} x_{j,t}^2} \tag{8}$$

or

$$\hat{\beta}_{ea} = \sum_{i=1,2} \sum_{j=1,2} \check{w}_{i,j} \hat{\beta}_{i.j} \tag{9}$$

where

$$\hat{\beta}_{ij} = \frac{\sum_{t=1}^{T} y_{i,t} x_{j,t}}{\sum_{t=1}^{T} x_{j,t}^2} \tag{10}$$

$$\breve{w}_{i,j} = \frac{\sum_{t=1}^{T} x_{j,t}^2}{\sum_{i=1,2}\sum_{j=1,2}\sum_{t=1}^{T} x_{j,t}^2} \tag{11}$$

with $\sum_{i=1,2}\sum_{j=1,2} \breve{w}_{i,j} = \sum_{i=1,2}\sum_{j=1,2} \frac{\sum_{t=1}^{T} x_{j,t}^2}{\sum_{i=1,2}\sum_{j=1,2}\sum_{t=1}^{T} x_{j,t}^2} = 1.$

The ex-ante model averaging or stacked OLS estimator of β is then equivalent to its ex-post counterpart, with weights determined according to the *relative variation* of the candidate regressors.

Moreover, consistent OLS estimation of σ^2 from the generic i,jth disjoint model yields

$$\tilde{\sigma}_{i,j}^2 = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_{i,j,t}^2}{T} \tag{12}$$

while the stacked estimator is

$$\begin{aligned} \tilde{\sigma}_{ea}^2 &= \frac{\sum_{i=1,2}\sum_{j=1,2}\sum_{t=1}^{T} \hat{\varepsilon}_{i,j,t}^2}{S} \\ &= \frac{1}{4}\sum_{i=1,2}\sum_{j=1,2} \frac{\sum_{t=1}^{T} \hat{\varepsilon}_{i,j,t}^2}{T} \\ &= \frac{1}{4}\sum_{i=1,2}\sum_{j=1,2} \tilde{\sigma}_{i,j}^2 \end{aligned} \tag{13}$$

Hence, the stacked OLS estimator of σ^2 is equivalent to the arithmetic mean, across models, of the disjoint OLS estimators of σ^2.

Issues related to the (relative) efficiency of the stacked OLS estimator and the gain in terms of higher degrees of freedom are discussed below.

3. Ex-ante model averaging by stacking

Consider the regression function

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{14}$$

and suppose that P candidate dependent variables are available, i.e., $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_P$, where $\mathbf{y}_p$, $p = 1, .., P$, is a $T \times 1$ column vector of observations.

For simplicity, three cases for the specification of the design matrix are considered:

1. The case of a single $T \times K$ design matrix $\mathbf{X}$ for the K regressors $\mathbf{x}_k$, $k = 1, ..., K$, where $\mathbf{x}_k$ is a $T \times 1$ vector and $T > K$.

2. The case of R candidates for *one* of the K regressors in the model, ordered first for simplicity, i.e., $\mathbf{x}_{1r}$, $r = 1, ..., R$, yielding up to R different $\mathbf{X}_r$ design matrices.

3. The case of R candidates for *each* of the K regressors in the model, yielding up to R^K different design matrices $\mathbf{X}_r$, $r = 1, .., R^K$.

3.1. The case of a single design matrix

In case 1. up to P models could be estimated, i.e.,

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \\ \mathbf{y}_2 &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \\ &... \\ \mathbf{y}_P &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_P \end{aligned} \tag{15}$$

Their union yields the stacked model

$$\mathbf{y_{P,1}} = \mathbf{X_{P,1}}\boldsymbol{\beta} + \boldsymbol{\varepsilon_{P,1}} \tag{16}$$

where $\mathbf{y_{P,1}} = \left[\begin{array}{cccc} \mathbf{y}_1' & \mathbf{y}_2' & ... & \mathbf{y}_P' \end{array} \right]'$ is a $(P \times T) \times 1$ vector of observations on the P available candidate dependent variables, obtained by stacking the P column vectors $\mathbf{y}_i$ on top of one other; $\mathbf{X_{P,1}}$ is the $(P \times T) \times K$ joint design matrix obtained by staking P times the matrix $\mathbf{X}$ on top of itself, i.e., $\mathbf{X_{P,1}} = \left[\begin{array}{cccc} \mathbf{X}' & \mathbf{X}' & ... & \mathbf{X}' \end{array} \right]'$, $\boldsymbol{\beta}$ is the $K \times 1$ vector of parameters, and $\boldsymbol{\varepsilon_{P,1}}$ is the $(P \times T) \times 1$ vector of residuals $\boldsymbol{\varepsilon_{P,1}} = \left[\begin{array}{cccc} \boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_2' & ... & \boldsymbol{\varepsilon}_P' \end{array} \right]'$, obtained by stacking the P column vectors $\boldsymbol{\varepsilon}_i$ on top of one other. Hence, the sample size of the stacked model is $S = P \times T$.

Disjoint OLS estimation of the pth generic model in (15) yields (see [9])

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_p \tag{17}$$

while for the variance, in large samples,

$$\tilde{\sigma}_p^2 = \frac{\hat{\boldsymbol{\varepsilon}}_p'\hat{\boldsymbol{\varepsilon}}_p}{T}. \tag{18}$$

3.1.1 The ex-ante model averaging estimator

Ex-ante model averaging is obtained by OLS estimation of the stacked model in (16), yielding

$$\hat{\boldsymbol{\beta}}_{ea} = (\mathbf{X}_{\mathbf{P,1}}'\mathbf{X}_{\mathbf{P,1}})^{-1}\mathbf{X}_{\mathbf{P,1}}'\mathbf{y}_{\mathbf{P,1}} \tag{19}$$

$$\tilde{\sigma}_{ea}^2 = \frac{\hat{\boldsymbol{\varepsilon}}_{\mathbf{P,1}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{P,1}}}{S}. \tag{20}$$

The linkage between ex-ante and ex-post model averaging can then be gauged by noting that (19) can be stated as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ea} &= \left[\mathbf{X}'\mathbf{X} + \mathbf{X}'\mathbf{X} + ... + \mathbf{X}'\mathbf{X}\right]^{-1} \left[\begin{array}{c} \mathbf{X}'\mathbf{y}_1 + \mathbf{X}'\mathbf{y}_2 + ... + \mathbf{X}'\mathbf{y}_P \end{array} \right] \\ &= [P\mathbf{X}'\mathbf{X}]^{-1} \left[\begin{array}{c} \mathbf{X}'\mathbf{y}_1 + \mathbf{X}'\mathbf{y}_2 + ... + \mathbf{X}'\mathbf{y}_P \end{array} \right] \\ &= \frac{1}{P}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_1 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_2 + ... + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_P\right] \\ &= \frac{1}{P}\sum_{p=1}^{P}\hat{\boldsymbol{\beta}}_p \end{aligned} \tag{21}$$

where $\hat{\boldsymbol{\beta}}_p = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_p$, $p = 1, ..., P$.

Hence, in this case, ex-ante OLS model averaging is equivalent to ex-post *arithmetic* model averaging across the P disjoint OLS estimators $\hat{\boldsymbol{\beta}}_p$.

Similarly for $\tilde{\sigma}^2_{ea}$

$$\begin{aligned}\tilde{\sigma}^2_{ea} &= \frac{\hat{\boldsymbol{\varepsilon}}'_{\mathbf{P},\mathbf{1}}\hat{\boldsymbol{\varepsilon}}_{\mathbf{P},\mathbf{1}}}{S} \\ &= \frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\boldsymbol{\varepsilon}}'_p\hat{\boldsymbol{\varepsilon}}_p}{T} \\ &= \frac{1}{P}\sum_{p=1}^{P}\tilde{\sigma}^2_p\end{aligned} \tag{22}$$

which also is the arithmetic average, across the P available models, of the disjoint estimators $\tilde{\sigma}^2_p$.

3.2. The case of multiple design matrices

In the case of multiple design matrices, up to G regression models can be computed, with $G = R$ in case 2. and $G = R^K$ in case 3., i.e.,

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{1,1} \\ \mathbf{y}_1 &= \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{1,2} \\ &\vdots \\ \mathbf{y}_1 &= \mathbf{X}_G\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{1,G} \\ \mathbf{y}_2 &= \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{2,1} \\ \mathbf{y}_2 &= \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{2,2} \\ &\vdots \\ \mathbf{y}_2 &= \mathbf{X}_G\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{2,G} \\ &\vdots \\ \mathbf{y}_P &= \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{P,1} \\ \mathbf{y}_P &= \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{P,2} \\ &\vdots \\ \mathbf{y}_P &= \mathbf{X}_G\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{P,G.}\end{aligned} \tag{23}$$

The disjoint OLS estimator for the generic p, rth model, $p = 1, ..., P$, $r = 1, ..., R$, in (23)

$$\mathbf{y}_p = \mathbf{X}_r\boldsymbol{\beta}_{p,r} + \boldsymbol{\varepsilon}_{p,r} \tag{24}$$

is

$$\hat{\boldsymbol{\beta}}_{p,r} = \left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1}\mathbf{X}'_r\mathbf{y}_p \tag{25}$$

while for the variance, in large samples,

$$\tilde{\sigma}^2_{p,r} = \frac{\hat{\boldsymbol{\varepsilon}}'_{p,r}\hat{\boldsymbol{\varepsilon}}_{p,r}}{T}. \tag{26}$$

On the other hand, the union of the above disjoint models yields the stacked model

$$\mathbf{y}_{\mathbf{P},\mathbf{G}} = \mathbf{X}_{\mathbf{P},\mathbf{G}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\mathbf{P},\mathbf{G}} \tag{27}$$

where $\boldsymbol{\beta}$ is the $K \times 1$ vector of parameters, $\mathbf{y}_{\mathbf{P},\mathbf{G}} = vec\left(\mathbf{i}_G \otimes \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & ... & \mathbf{y}_P \end{bmatrix}\right)$ is the $(T \times P \times G) \times 1$ vector collecting the P $\mathbf{y}_p$ $(T \times 1)$ vectors, $p = 1, .., P$, which are then stacked on top of one other G times, *vec* is the vectorization operator, $\otimes$ is the Kronecker product and $\mathbf{i}_G$ a $G \times 1$ unitary vector.[3]

[3] Hence, $\mathbf{y}_{\mathbf{P},\mathbf{G}} = \Big[\underset{1\times(T\times G)}{\begin{pmatrix} \mathbf{y}'_1 & \mathbf{y}'_1 & ... & \mathbf{y}'_1 \end{pmatrix}}\underset{1\times(T\times G)}{\begin{pmatrix} \mathbf{y}'_2 & \mathbf{y}'_2 & ... & \mathbf{y}'_2 \end{pmatrix}}...\underset{1\times(T\times G)}{\begin{pmatrix} \mathbf{y}'_P & \mathbf{y}'_P & ... & \mathbf{y}'_P \end{pmatrix}}\Big]'$.

By denoting $\mathbf{X}_* = \left[\begin{array}{cccc} \mathbf{X}_1^{'} & \mathbf{X}_2^{'} & ... & \mathbf{X}_G^{'} \end{array} \right]^{'}$ the $(G \times T) \times K$ matrix obtained by stacking the G candidate design matrices on top of one another, $\mathbf{X_{P,G}}$ is then the $(P \times G \times T) \times K$ design matrix yield by staking P times the matrix $\mathbf{X}_*$ on top of itself, i.e., $\mathbf{X_{P,G}} = \left[\begin{array}{cccc} \mathbf{X}_*^{'} & \mathbf{X}_*^{'} & ... & \mathbf{X}_*^{'} \end{array} \right]^{'}$. Finally, $\boldsymbol{\varepsilon}_{\mathbf{P,G}} = \left[\begin{array}{ccccccc} \boldsymbol{\varepsilon}_{1,1}^{\prime} & ... & \boldsymbol{\varepsilon}_{1,G}^{\prime} & ... & \boldsymbol{\varepsilon}_{P,1}^{\prime} & ... & \boldsymbol{\varepsilon}_{P,G}^{\prime} \end{array} \right]^{'}$ is a $(P \times G \times T) \times 1$ vector of residuals. Hence, the sample size of the stacked model is $S = T \times P \times G$.

The stacked OLS estimator is then computed as

$$\hat{\boldsymbol{\beta}}_{ea} = \left(\mathbf{X}_{\mathbf{P,G}}^{\prime}\mathbf{X}_{\mathbf{P,G}}\right)^{-1}\mathbf{X}_{\mathbf{P,G}}^{\prime}\mathbf{y}_{\mathbf{P,G}} \tag{28}$$

$$\tilde{\sigma}_{ea}^2 = \frac{\hat{\boldsymbol{\varepsilon}}_{\mathbf{P,G}}^{\prime}\hat{\boldsymbol{\varepsilon}}_{\mathbf{P,G}}}{S}. \tag{29}$$

3.2.1 The case of a single candidate dependent variable

For sake of simplicity, consider first the case where $P = 1$; hence, $S = G \times T$, $\mathbf{y_{P,G}} = \mathbf{y_{1,G}} = \mathbf{i}_G \otimes \mathbf{y}_1$, and the design matrix in the stacked model is $\mathbf{X_{P,G}} = \mathbf{X_{1,G}} = \mathbf{X}_* = \left[\begin{array}{cccc} \mathbf{X}_1^{'} & \mathbf{X}_2^{'} & ... & \mathbf{X}_G^{'} \end{array} \right]^{'}$.

The stacked OLS estimator in (28) can then be stated

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ea} &= \left[\mathbf{X}_*^{'}\mathbf{X}_*\right]^{-1}\left[\begin{array}{c} \mathbf{X}_*^{'}\mathbf{y_{1,G}} \end{array} \right] \\ &= \left[\mathbf{X}_*^{'}\mathbf{X}_*\right]^{-1} \times \left[\begin{array}{c} \mathbf{X}_1^{\prime}\mathbf{y}_1 + \mathbf{X}_2^{\prime}\mathbf{y}_1 + ... + \mathbf{X}_G^{\prime}\mathbf{y}_1 \end{array} \right] \\ &= \left[\mathbf{X}_*^{'}\mathbf{X}_*\right]^{-1}\mathbf{X}_1^{\prime}\mathbf{y}_1 + \left[\mathbf{X}_*^{'}\mathbf{X}_*\right]^{-1}\mathbf{X}_2^{\prime}\mathbf{y}_1 + ... + \left[\mathbf{X}_*^{'}\mathbf{X}_*\right]^{-1}\mathbf{X}_G^{\prime}\mathbf{y}_1 \end{aligned} \tag{30}$$

where $\mathbf{X}_*^{'}\mathbf{X}_* = \mathbf{X}_1^{\prime}\mathbf{X}_1 + \mathbf{X}_{\mathbf{2}}^{\prime}\mathbf{X_2} + ... + \mathbf{X}_G^{\prime}\mathbf{X}_G$.

Denote $\mathbf{K}_r = \sum_{i=1,\ i\neq r}^{G} \mathbf{X}_i^{\prime}\mathbf{X}_i$, yielding $\mathbf{K}_1 = \mathbf{X}_2^{\prime}\mathbf{X}_2 + \mathbf{X}_3^{\prime}\mathbf{X}_3 + ... + \mathbf{X}_G^{\prime}\mathbf{X}_G$, $\mathbf{K}_2 = \mathbf{X}_1^{\prime}\mathbf{X}_1 + \mathbf{X}_3^{\prime}\mathbf{X}_3 + ... + \mathbf{X}_G^{\prime}\mathbf{X}_G$, and so on. By substitution in (30), it follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ea} &= \left(\left[\mathbf{X}_1^{\prime}\mathbf{X}_1 + \mathbf{K}_1\right]^{-1}\mathbf{X}_1^{\prime}\mathbf{y}_1\right) + \left(\left[\mathbf{X}_2^{\prime}\mathbf{X}_2 + \mathbf{K}_2\right]^{-1}\mathbf{X}_2^{\prime}\mathbf{y}_1\right) + ... + \left(\left[\mathbf{X}_G^{\prime}\mathbf{X}_G + \mathbf{K}_G\right]^{-1}\mathbf{X}_G^{\prime}\mathbf{y}_1\right) \\ &= \sum_{r=1}^{G}\left[\mathbf{X}_r^{\prime}\mathbf{X}_r + \mathbf{K}_r\right]^{-1}\mathbf{X}_r^{\prime}\mathbf{y}_1. \end{aligned} \tag{31}$$

Using matrix inversion rules[4], one has

$$\begin{aligned} \left[\mathbf{X}_r^{\prime}\mathbf{X}_r + \mathbf{K}_r\right]^{-1} &= \left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1} - \left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1}\left(\mathbf{K}_r^{-1} + \left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1}\right)^{-1}\left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1} \\ &= \left(\mathbf{I}_K - \mathbf{K}_r^{*}\right)\left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1}. \end{aligned} \tag{32}$$

where $\mathbf{K}_r^{*} = \left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1}\left(\mathbf{K}_r^{-1} + \left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1}\right)^{-1}$.

By substitution in (31), it follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ea} &= \sum_{r=1}^{G}\left(\mathbf{I}_K - \mathbf{K}_r^{*}\right)\left(\mathbf{X}_r^{\prime}\mathbf{X}_r\right)^{-1}\mathbf{X}_r^{\prime}\mathbf{y}_1 \\ &= \sum_{r=1}^{G}\left(\mathbf{I}_K - \mathbf{K}_r^{*}\right)\hat{\boldsymbol{\beta}}_{1,r} \\ &= \sum_{r=1}^{G}\check{\mathbf{W}}_r^{*}\hat{\boldsymbol{\beta}}_{1,r}. \end{aligned} \tag{33}$$

[4] Given matrices A and C, non singular and of proper dimensions for their sum, it holds $(A+C)^{-1} = A^{-1}(C^{-1}+A^{-1})^{-1}A^{-1}$.

where $\hat{\boldsymbol{\beta}}_{1,r} = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{y}_1$.

Optimal ex-ante weights, contained in the $K \times K$ matrices $\check{\mathbf{W}}_r^*$, $r = 1, ..., G$, are then computed by taking into account all the information available on the various candidate regressors, being proportional to their relative variation. In fact, multiplying both sides of (32) by $\mathbf{X}_r'\mathbf{X}_r$, one has

$$[\mathbf{X}_r'\mathbf{X}_r + \mathbf{K}_r]^{-1}(\mathbf{X}_r'\mathbf{X}_r) = (\mathbf{I}_K - \mathbf{K}_r^*)$$

and therefore $\sum_{r=1}^{G} \check{\mathbf{W}}_r^* = \sum_{r=1}^{G} [\mathbf{X}_r'\mathbf{X}_r + \mathbf{K}_r]^{-1}(\mathbf{X}_r'\mathbf{X}_r) = \mathbf{I}_K$.

Moreover, given $\hat{\boldsymbol{\varepsilon}}_{\mathbf{P},\mathbf{G}} = \hat{\boldsymbol{\varepsilon}}_{\mathbf{1},\mathbf{G}}$, one has

$$\begin{aligned}
\tilde{\sigma}_{ea}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}_{\mathbf{1},\mathbf{G}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{1},\mathbf{G}}}{S} \\
&= \frac{1}{G}\sum_{r=1}^{G}\frac{\hat{\boldsymbol{\varepsilon}}_{1,r}'\hat{\boldsymbol{\varepsilon}}_{1,r}}{T} \\
&= \frac{1}{G}\sum_{r=1}^{G}\tilde{\sigma}_{1,r}^2.
\end{aligned}$$

Hence, $\tilde{\sigma}_{ea}^2$ is the arithmetic average, across the available G models, of the disjoint estimators $\tilde{\sigma}_{1,r}^2$.

3.2.2 The case of multiple candidate dependent variable

Consider now the case in which more than single candidate dependent variable is available, i.e., $P > 1$. The stacked OLS estimator in (28) is then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ea} &= \left[\mathbf{X}_*'\mathbf{X}_* + \mathbf{X}_*'\mathbf{X}_* + ... + \mathbf{X}_*'\mathbf{X}_*\right]^{-1}\left[\ \mathbf{X}_*'\mathbf{y}_1 + \mathbf{X}_*'\mathbf{y}_2 + ... + \mathbf{X}_*'\mathbf{y}_P\ \right] \qquad (34)\\
&= \left[P\left(\mathbf{X}_*'\mathbf{X}_*\right)\right]^{-1} \times [(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_1 + ... + \mathbf{X}_G'\mathbf{y}_1) \\
&\quad + (\mathbf{X}_1'\mathbf{y}_2 + \mathbf{X}_2'\mathbf{y}_2 + ... + \mathbf{X}_G'\mathbf{y}_2) + ... + (\mathbf{X}_1'\mathbf{y}_P + \mathbf{X}_2'\mathbf{y}_P + ... + \mathbf{X}_G'\mathbf{y}_P)] \\
&= \frac{1}{P}\left[\mathbf{X}_*'\mathbf{X}_*\right]^{-1} \times [(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{X}_2'\mathbf{y}_1 + ... + \mathbf{X}_G'\mathbf{y}_1) + (\mathbf{X}_1'\mathbf{y}_2 + \mathbf{X}_2'\mathbf{y}_2 + ... + \mathbf{X}_G'\mathbf{y}_2) + \\
&\quad ... + (\mathbf{X}_1'\mathbf{y}_P + \mathbf{X}_2'\mathbf{y}_P + ... + \mathbf{X}_G'\mathbf{y}_P)] \\
&= \frac{1}{P}\left[\mathbf{X}_*'\mathbf{X}_*\right]^{-1} \times \left[\sum_{p=1}^{P}\mathbf{X}_1'\mathbf{y}_p + \sum_{p=1}^{P}\mathbf{X}_2'\mathbf{y}_p + ... + \sum_{p=1}^{P}\mathbf{X}_G'\mathbf{y}_p\right]
\end{aligned}$$

where again $\mathbf{X}_*'\mathbf{X}_* = \mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_\mathbf{2}'\mathbf{X}_\mathbf{2} + ... + \mathbf{X}_G'\mathbf{X}_G$.

Moreover, denote $\mathbf{K}_r = \sum_{i=1,\ i\neq r}^{G}\mathbf{X}_i'\mathbf{X}_i$, i.e., $\mathbf{K}_1 = \mathbf{X}_2'\mathbf{X}_2 + \mathbf{X}_3'\mathbf{X}_3 ... + \mathbf{X}_G'\mathbf{X}_G$, $\mathbf{K}_2 = \mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_3'\mathbf{X}_3 ... + \mathbf{X}_G'\mathbf{X}_G$, and so on, by substitution in (34), one then has

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ea} &= \left(\frac{1}{P}[\mathbf{X}_1'\mathbf{X}_1 + \mathbf{K}_1]^{-1}\sum_{p=1}^{P}\mathbf{X}_1'\mathbf{y}_p\right) + \left(\frac{1}{P}[\mathbf{X}_2'\mathbf{X}_2 + \mathbf{K}_2]^{-1}\sum_{p=1}^{P}\mathbf{X}_2'\mathbf{y}_p\right) \qquad (35)\\
&\quad + ... + \left(\frac{1}{P}[\mathbf{X}_G'\mathbf{X}_G + \mathbf{K}_G]^{-1}\sum_{p=1}^{P}\mathbf{X}_G'\mathbf{y}_p\right) \\
&= \frac{1}{P}\sum_{r=1}^{G}[\mathbf{X}_r'\mathbf{X}_r + \mathbf{K}_r]^{-1}\sum_{p=1}^{P}\mathbf{X}_r'\mathbf{y}_p.
\end{aligned}$$

By recalling that $[\mathbf{X}_r'\mathbf{X}_r+\mathbf{K}_r]^{-1} = (\mathbf{I}_K - \mathbf{K}_r^*)(\mathbf{X}_r'\mathbf{X}_r)^{-1}$, where $\mathbf{K}_r^* = (\mathbf{X}_r'\mathbf{X}_r)^{-1}(\mathbf{K}_r^{-1} + (\mathbf{X}_r'\mathbf{X}_r)^{-1})^{-1}$, by substitution in (35) one eventually has

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ea} &= \frac{1}{P}\sum_{r=1}^{G}(\mathbf{I}_K - \mathbf{K}_r^*)\sum_{p=1}^{P}(\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{y}_p \\ &= \sum_{r=1}^{G}(\mathbf{I}_K - \mathbf{K}_r^*)\left(\frac{1}{P}\sum_{p=1}^{P}\hat{\boldsymbol{\beta}}_{p,r}\right) \\ &= \sum_{r=1}^{G}\breve{\mathbf{W}}_r^*\left(\frac{1}{P}\sum_{p=1}^{P}\hat{\boldsymbol{\beta}}_{p,r}\right)\end{aligned} \tag{36}$$

where, as for the previous case, $\sum_{r=1}^{G}\breve{\mathbf{W}}_r^* = \sum_{r=1}^{G}[\mathbf{X}_r'\mathbf{X}_r+\mathbf{K}_r]^{-1}(\mathbf{X}_r'\mathbf{X}_r) = \mathbf{I}_K$.

The optimal ex-ante weights, contained in the $K \times K$ matrices $\breve{\mathbf{W}}_r^*$, $r = 1, ..., G$, are again computed by taking into account all the information available on the various candidate regressors and are proportional to their relative variation. Averaging is then performed across all possible models which can be estimated according to the P candidate dependent variables.

Moreover,

$$\begin{aligned}\tilde{\sigma}_{ea}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}_{\mathbf{P},\mathbf{G}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{P},\mathbf{G}}}{S} \\ &= \frac{1}{G}\sum_{r=1}^{G}\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\boldsymbol{\varepsilon}}_{p,r}'\hat{\boldsymbol{\varepsilon}}_{p,r}}{T} \\ &= \frac{1}{G}\sum_{r=1}^{G}\frac{1}{P}\sum_{p=1}^{P}\tilde{\sigma}_{p,r}^2.\end{aligned} \tag{37}$$

Then, ex-ante model averaging estimation of the variance $\tilde{\sigma}_{ea}^2$ is computed as the arithmetic average, across all the $G \times P$ models, of the disjoint estimators $\tilde{\sigma}_{p,r}^2$.

4. Statistical properties

Assume that the properties of the classical linear regression model hold, i.e.:

1. The population regression function is linear in the K parameters, i.e., $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
2. $\{y_{p,t}, \mathbf{x}_{r,t}\}$ is a candidate stationary and ergodic process, $p = 1, .., P$; $r = 1, ..., G$, $G = R, R^K$; where $\mathbf{x}_{r,t}$ is a $K \times 1$ vector of regressors (belonging to the rth design matrix $\mathbf{X}_r$) at observation t, $t = 1, ..., T$; $T > K$.
3. The regressors $\mathbf{x}_{r,t}$ are at least contemporaneously orthogonal to the residuals, i.e., $E[\varepsilon_{p,r,t}|\mathbf{x}_{r,t}] = 0$, where $\varepsilon_{p,r,t}$ is the residual from the generic prth model at observation t.
4. Any of the $T \times K$ design matrices $\mathbf{X}_r$ has rank equal to K with probability 1, with $\text{plim}\big(T^{-1}\mathbf{X}_r'\mathbf{X}_r\big)$ a finite, symmetric, invertible, positive semidefinite matrix.
5. The conditional variance covariance matrix of the residuals $\varepsilon_{p,r,t}$ is a scalar identity matrix, i.e., $E\left[\boldsymbol{\varepsilon}_{p,r}\boldsymbol{\varepsilon}_{p,r}'|\mathbf{X}_r\right] \equiv \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, implying that the residuals are conditionally homoskedastic ($E\left[\varepsilon_{p,r,t}^2|\mathbf{x}_{r,t}\right] = \sigma^2$) and non serially correlated ($E[\varepsilon_{p,r,t}\varepsilon_{p,r,t-n}|\mathbf{x}_{r,t}] = 0$, $n = 1, ...$).

Under the above assumptions (even relaxing the conditional homoskedasticity property), the disjoint OLS estimator $\hat{\boldsymbol{\beta}}_{p,r}$ in (25) and $\hat{\boldsymbol{\sigma}}^2_{p,r}$ in (26) is consistent and asymptotically normal (see [9]). The same properties hold for the stacked OLS estimator. Proofs for the most general case are reported below; results for the intermediate cases can be straightforwardly derived from those provided, by setting $P = 1$ or $G = 1$.

4.1. Large sample properties

In so far as $\text{plim}\left(\hat{\boldsymbol{\beta}}_{p,r}\right) = \boldsymbol{\beta}$, it follows for $\hat{\boldsymbol{\beta}}_{ea}$ in (36)

$$\begin{aligned}
\text{plim}\left(\hat{\boldsymbol{\beta}}_{ea}\right) &= \text{plim}\left(\sum_{r=1}^{G} \check{\mathbf{W}}^*_r \left(\frac{1}{P}\sum_{p=1}^{P} \hat{\boldsymbol{\beta}}_{p,r}\right)\right) \\
&= \sum_{r=1}^{G} \text{plim}\left(\check{\mathbf{W}}^*_r\right) \times \text{plim}\left(\frac{1}{P}\sum_{p=1}^{P} \hat{\boldsymbol{\beta}}_{p,r}\right) \\
&= \sum_{r=1}^{G} \mathbf{V}^*_r \times \left(\frac{1}{P}\sum_{p=1}^{P} \text{plim}\ \left(\hat{\boldsymbol{\beta}}_{p,r}\right)\right) \\
&= \sum_{r=1}^{G} \mathbf{V}^*_r \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{\beta} \\
&= \sum_{\mathbf{r=1}}^{\mathbf{G}} \mathbf{V}^*_{\mathbf{r}} \boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

since by ergodic stationarity $\text{plim}\left(\check{\mathbf{W}}^*_r\right) = \mathbf{V}^*_r$, where $\mathbf{V}^*_r$ is a finite and non singular $K \times K$ matrix and $\sum_{r=1}^{G} \mathbf{V}^*_r = \mathbf{I}_K$.

Moreover, in so far as $\text{plim}\left(\tilde{\sigma}^2_{p,r}\right) = \sigma^2$, it follows for $\tilde{\sigma}^2_{ea}$ in (37)

$$\begin{aligned}
\text{plim}\left(\tilde{\sigma}^2_{ea}\right) &= \text{plim}\left(\frac{1}{G}\sum_{r=1}^{G}\frac{1}{P}\sum_{p=1}^{P} \tilde{\sigma}^2_{p,r}\right) \\
&= \frac{1}{G}\sum_{r=1}^{G}\frac{1}{P}\sum_{p=1}^{P} \text{plim}\left(\tilde{\sigma}^2_{p,r}\right) \\
&= \frac{1}{G}\sum_{r=1}^{G}\frac{1}{P}\sum_{p=1}^{P} \sigma^2 = \sigma^2.
\end{aligned}$$

Under properties 1. to 5., by means of a CLT (see [9]), one also has

$$S^{-1/2}\mathbf{X}'_{\mathbf{P,G}}\boldsymbol{\varepsilon}_{\mathbf{P,G}} \xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right)\right)$$

leading to

$$\begin{aligned}
\sqrt{S}\left(\hat{\boldsymbol{\beta}}_{ea} - \boldsymbol{\beta}\right) &\xrightarrow{d} \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right)^{-1} \times N\left(\mathbf{0}, \text{plim}\left(S^{-1}\sigma^2\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right)\right) \\
&\xrightarrow{d} N\left(\mathbf{0}, \sigma^2\left(\text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right)^{-1} \times \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right) \times \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right)^{-1}\right)\right) \\
&\xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P,G}}\mathbf{X}_{\mathbf{P,G}}\right)^{-1}\right).
\end{aligned}$$

The asymptotic distribution of $\hat{\boldsymbol{\beta}}_{ea}$ then follows

$$\hat{\boldsymbol{\beta}}_{ea} \overset{asy}{\sim} N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right)$$

as well as its feasible form

$$\hat{\boldsymbol{\beta}}_{ea} \overset{asy}{\sim} N\left(\boldsymbol{\beta}, \tilde{\sigma}^2_{ea} \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right).$$

In the case of conditional heteroskedasticity ($\Sigma = Diag(\sigma_t^2)$), it would be straightforward to prove that

$$\begin{aligned}\sqrt{S}\left(\hat{\boldsymbol{\beta}}_{ea} - \boldsymbol{\beta}\right) &\xrightarrow{d} \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1} \times N\left(\mathbf{0}, \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}\boldsymbol{\Sigma}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)\right) \\ &\xrightarrow{d} N\left(\mathbf{0}, \left(\text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1} \times \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}\boldsymbol{\Sigma}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right) \times \text{plim}\left(S^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right)\right)\end{aligned}$$

and

$$\hat{\boldsymbol{\beta}}_{ea} \overset{asy}{\sim} N\left(\boldsymbol{\beta}, \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\boldsymbol{\Sigma}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right)$$

with feasible form

$$\hat{\boldsymbol{\beta}}_{ea} \overset{asy}{\sim} N\left(\boldsymbol{\beta}, \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\hat{\boldsymbol{\Sigma}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right)$$

where $\hat{\Sigma} = Diag(\hat{\sigma}_t^2)$.

The relative efficiency of the stacked over the disjoint OLS estimator can be established by comparing their asymptotic variances, i.e. $asyV\left[\hat{\boldsymbol{\beta}}_{p,r}\right]$ and $asyV\left[\hat{\boldsymbol{\beta}}_{ea}\right]$. One then has

$$\begin{aligned}asyV\left[\hat{\boldsymbol{\beta}}_{p,r}\right] - asyV\left[\hat{\boldsymbol{\beta}}_{ea}\right] &= \sigma^2\left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1} - \sigma^2\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1} - \sigma^2\frac{1}{P}\left(\mathbf{I}_K - \mathbf{K}^*_r\right)\left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1} \\ &= \sigma^2\frac{P-1}{P}\left(\mathbf{I} + \mathbf{K}^*_r\right)\left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1} \qquad (38)\end{aligned}$$

which is a finite, symmetric, positive semidefinite $K \times K$ matrix, as $\sigma^2 > 0$ and $\frac{P-1}{P} > 0$, both finite, and $\left(\mathbf{I} + \mathbf{K}^*_r\right)\left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1}$ is a finite, symmetric, positive semidefinite $K \times K$ matrix by construction ($\mathbf{X}_r$ is real and of full column rank $K < T$ for any r).

Finally, the gain in terms of degrees of freedom yield by the stacked over the disjoint OLS estimator is equal to $(P \times G - 1) \times T$. In fact, by recalling that the number of degrees of freedom of the residual term is equal to the rank of the annihilator matrix (see [9]), the gain yield by stacked over disjoint OLS estimation can be established by comparing the rank of the annihilator matrix in the two cases

$$\mathbf{M}_{ea} = \mathbf{I}_{P\times G\times T} - \mathbf{X}_{\mathbf{P},\mathbf{G}}\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}$$

which is of rank equal to $P \times G \times T - K$ as

$$\begin{aligned}rank(\mathbf{M}_{ea}) &= trace(\mathbf{I}_{P\times G\times T}) - trace\left(\mathbf{X}_{\mathbf{P},\mathbf{G}}\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\mathbf{X}'_{\mathbf{P},\mathbf{G}}\right) \\ &= P \times G \times T - trace(\mathbf{I}_K) = P \times G \times T - K\end{aligned}$$

and

$$\mathbf{M}_{p,r} = \mathbf{I}_T - \mathbf{X}_r\left(\mathbf{X}'_r\mathbf{X}_r\right)^{-1}\mathbf{X}'_r$$

which is of rank $T - K$ as

$$\begin{aligned} rank(\mathbf{M}_{p,r}) &= trace(\mathbf{I}_T) - trace\left(\mathbf{X}_r \left(\mathbf{X}_r'\mathbf{X}_r\right)^{-1} \mathbf{X}_r'\right) \\ &= T - trace(\mathbf{I}_K) = T - K. \end{aligned}$$

The increase in degrees of freedom yield by stacked over disjoint OLS estimation is then $(P \times G \times T - K) - (T - K) = (P \times G - 1) \times T$.

4.2. Small sample properties

If the stronger assumption of strict exogeneity is made in 3. above, i.e., $E\left[\varepsilon_{p,r,t}|\mathbf{X}_r\right] = 0$, the disjoint OLS estimators $\hat{\boldsymbol{\beta}}_{p,r}$ in (25) and $\hat{\boldsymbol{\sigma}}^2_{p,r} = \dfrac{\hat{\boldsymbol{\varepsilon}}'_{p,r}\hat{\boldsymbol{\varepsilon}}_{p,r}}{T-k}$ are also (conditionally and unconditionally) $BLUE$, i.e., best unbiased and efficient (within the class of linear estimators) (see [9])[5]. Moreover, if the assumption of conditional Normality of the error term is added, i.e., $\hat{\boldsymbol{\varepsilon}}_{p,r}|\mathbf{X}_* \sim N\left(\mathbf{0}, \sigma^2\right)$, OLS is (conditionally and unconditionally) BUE, i.e., best unbiased (within the class of linear and non linear estimators), as well as (conditionally and unconditionally) Normally distributed

$$\hat{\boldsymbol{\beta}}_{p.r}|\mathbf{X}_{p,r} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'_{p,r}\mathbf{X}_{p,r}\right)^{-1}\right) \tag{39}$$

$$\hat{\boldsymbol{\beta}}_{p.r} \sim N\left(\boldsymbol{\beta}, \sigma^2 E\left[\left(\mathbf{X}'_{p,r}\mathbf{X}_{p,r}\right)^{-1}\right]\right)$$

where $E\left[\mathbf{X}'_{p,r}\mathbf{X}_{p,r}\right]$ is a finite, nonsingular, symmetric, positive semidefinite matrix of rank $K < T$.

The above properties can also be established for the stacked OLS estimator, in the same way as for the disjoint OLS estimator (see [9]), yielding

$$\hat{\boldsymbol{\beta}}_{ea}|\mathbf{X}_{\mathbf{P},\mathbf{G}} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right) \tag{40}$$

$$\hat{\boldsymbol{\beta}}_{ea} \sim N\left(\boldsymbol{\beta}, \sigma^2 E\left[\left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right]\right)$$

with $E\left[\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right]$ a finite, nonsingular, symmetric, positive semidefinite matrix of rank $K < T$, and feasible form

$$\hat{\boldsymbol{\beta}}_{ea} \sim N\left(\boldsymbol{\beta}, \hat{\sigma}^2_{ea} \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1}\right)$$

where $\hat{\sigma}^2_{ea} = \dfrac{\hat{\boldsymbol{\varepsilon}}'_{\mathbf{P},\mathbf{G}}\hat{\boldsymbol{\varepsilon}}_{\mathbf{P},\mathbf{G}}}{S-K}$.

Then, by comparing the conditional variances of $\hat{\boldsymbol{\beta}}_{p,r}$ and $\hat{\boldsymbol{\beta}}_{ea}$, one has again

$$\begin{aligned} V\left[\hat{\boldsymbol{\beta}}_{p,r}|\mathbf{X}_r\right] - V\left[\hat{\boldsymbol{\beta}}_{ea}|\mathbf{X}_{\mathbf{P},\mathbf{G}}\right] &= \sigma^2 \left(\mathbf{X}_r'\mathbf{X}_r\right)^{-1} - \sigma^2 \left(\mathbf{X}'_{\mathbf{P},\mathbf{G}}\mathbf{X}_{\mathbf{P},\mathbf{G}}\right)^{-1} \\ &= \sigma^2 \frac{P-1}{P} \left(\mathbf{I} + \mathbf{K}_r^*\right) \left(\mathbf{X}_r'\mathbf{X}_r\right)^{-1} \end{aligned} \tag{41}$$

as for the asymptotic case. Moreover,

$$\begin{aligned} V\left[\hat{\boldsymbol{\beta}}_{p,r}\right] - V\left[\hat{\boldsymbol{\beta}}_{ea}\right] &= E\left[V\left[\hat{\boldsymbol{\beta}}_{p,r}|\mathbf{X}_r\right]\right] - E\left[V\left[\hat{\boldsymbol{\beta}}_{ea}|\mathbf{X}_{\mathbf{P},\mathbf{G}}\right]\right] \\ &= \sigma^2 \frac{P-1}{P} E\left[\left(\mathbf{I} + \mathbf{K}_r^*\right) \left(\mathbf{X}_r'\mathbf{X}_r\right)^{-1}\right] \end{aligned} \tag{42}$$

[5] The usual caveat concerning the efficiency of $\hat{\boldsymbol{\sigma}}^2_{p,r}$ applies, as no linear unbiased estimator of σ^2 achieves the Cramer-Rao Lower Bound, which is obtained by the biased ML estimator $\tilde{\boldsymbol{\sigma}}^2_{p,r}$.

which similarly is a finite, symmetric and positive semidefinite $K \times K$ matrix by construction.

Finally, the gain in terms of degrees of freedom yield by stacked over disjoint OLS estimation is again $(P \times G \times T - K) - (T - K) = (P \times G - 1) \times T$, as already shown for the asymptotic case.

5. Conclusions

The paper introduces an ex-ante model averaging approach, requiring the estimation of a single augmented model obtained from the union of all the possible candidate models, rather than their disjoint estimation. In this framework, optimal weights are implicitly computed according to the MSE metric, i.e., by minimizing the squared Euclidean distance between actual and predicted value vectors, and are proportional to the relative variation of the regressors. By exploiting ex-ante all the available information on the various candidate set of variables, and relying on more degrees of freedom, it then leads to more accurate and (relatively) more efficient estimation than available ex-post methods. Moreover, the proposed estimator shows the same optimal properties of the disjoint OLS estimator, under the usual set of assumptions concerning the population regression model. While the method is proposed to be used within the OLS estimator framework, extension to GIVE and GMM is straightforward. We point to [1] for an empirical application, using OLS and GMM estimation.

Acknowledgments

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 3202782013-2015. The flowers are supported by the branches/The trunk supports the branches/The roots support the trunk/But we do not see the roots (Mitsuo Aida).

References

[1] D. Baiardi and C. Morana. The financial Kuznets curve: Financial deepening and inequality in the euro area. *University of Milan-Bicocca, mimeo.*

[2] G. Claeskens and N. Hjort. *Model Selection and Model Averaging.* Cambridge University Press, Cambridge (UK), 2008. ISBN 9780521852258.

[3] D. F. Hendry and J. A. Doornik. *Empirical Model Discovery and Theory Evaluation.* MIT Press, Cambridge (US), 2004. ISBN 9780262028356.

[4] J. A. Hoeting, D. Madigan, A. E. Raftery and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999. MR 2001a:62033.

[5] E. Moral Benito. Model averaging in Economics: An overview. *Journal of Economic Surveys*, 29:46–75, 2015. DOI: 10.1111/joes.12044.

[6] S. Buckland, K. Burnham and N. Augustin. Model selection: An integral part of inference. *Biometrics*, 53:603–618. DOI 10.2307/2533961.

[7] B. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1180, 2007. DOI 10.1111/j.1468-0262.2007.00785.x.

[8] B. Hansen and J. Rancine. Jacknife model averaging. *Journal of Econometrics*, 167:38–46, 2010. DOI 10.1016/j.jeconom.2011.06.019.

[9] F. Hayashi. *Econometrics.* Princeton University Press, Princeton (US), 2000. ISBN 9780691010182.