Economics Letters 142 (2016) 69–73

Contents lists available at ScienceDirect

Economics Letters

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

On the dominance of Mallows model averaging estimator over ordinary least squares estimator

Xinyu Zhang [a], Aman Ullah [b], Shangwei Zhao [c,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[b] Department of Economics, University of California, Riverside, CA, USA
[c] College of Science, Minzu University of China, Beijing, China

ARTICLE INFO

Article history:
Received 18 February 2016
Accepted 23 February 2016
Available online 7 March 2016

JEL classification:
C13
C21

Keywords:
Finite sample size
Mean squared error
Model averaging
Sufficient condition

ABSTRACT

This note studies Mallows model averaging method for finite sample size situation. Sufficient conditions under which the model averaging estimator dominates the ordinary least squares estimator are provided with respect to mean squared error.

1. Introduction

Over the past two decades, there has been a substantial amount of interest in model averaging. Within the Bayesian paradigm, model averaging has long been a popular approach; see, for example, Hoeting et al. (1999) for a comprehensive review. In recent years, within the frequentest paradigm, model averaging methods have been proposed, including weighting strategies using scores of information criteria (Buckland et al., 1997; Hjort and Claeskens, 2003; Claeskens et al., 2006; Zhang and Liang, 2011; Zhang et al., 2012), asymptotically optimal methods (Hansen, 2007; Liang et al., 2011; Hansen and Racine, 2012; Liu and Okui, 2013), model averaging marginal regression (Li et al., 2015), among others. Frequentist model averaging technique has also been utilized in many contexts such as constructing optimal instruments (Kuersteiner and Okui, 2010), autoregressive models (Hansen, 2010), mixed-effects models (Zhang et al., 2014), and quantile regression models (Lu and Su, 2015).

However, in most of the existing literature, large sample properties on model averaging methods are mainly focused on and model averaging methods were generally compared asymptotically or by simulation performance (see, for example Liang et al., 2011, Hansen, 2014 and Liu, 2015). As far as we know, there was no analytical finite sample study on the condition under which a model averaging estimator dominates the ordinary least squares (OLS) estimator with respect to mean squared error (MSE). The main contribution of this paper is to develop results on the exact dominance of a model average estimator over the OLS estimator. The results are exact in the sense that they are valid for any sample size, especially when the sample size is small.

The remainder of this paper is organized as follows. Section 2 introduces some estimators and basic theoretic results. Section 3 presents the MSE comparison between the model average estimators and the OLS estimator, and provides the sufficient conditions under which these model averaging estimators dominate the OLS estimator. Section 4 concludes the paper. Technical proofs are contained in Appendix.

2. Estimation

We are concerned with a linear regression model:

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i, \quad e_i \sim \mathrm{Normal}(0, \sigma^2), \ i = 1, \ldots, n \tag{2.1}$$

where y_i is a scalar dependent variable, $\mathbf{x}_i(p \times 1)$ are independent variables, $\boldsymbol{\beta}(p \times 1)$ is a coefficient vector, e_i is an error term, and $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$ are assumed to be independent. Following

* Corresponding author.
E-mail address: zhaoshangwei@amss.ac.cn (S. Zhao).

http://dx.doi.org/10.1016/j.econlet.2016.02.027
0165-1765/

classic literature on finite sample studies such as James and Stein (1961) and Stein (1981), here we also assume normal distribution. To simplify notation we treat the independent variables as fixed, but the theory applies also to random independent variables if proper conditions are imposed. In matrix notation, the model (2.1) can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, $\mathbf{e} = (e_1, \ldots, e_n)^{\mathrm{T}} \sim \mathrm{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. We assume that $\mathbf{X}$ has full column rank $p < n$. The OLS estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \sim \mathrm{Normal}\left\{\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\}. \tag{2.3}$$

The variance σ^2 is estimated by $\widehat{\sigma}^2 = (n-p)^{-1}\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{y}\|^2$, where $\|\cdot\|^2$ stands for the Euclidean norm. It is well known that $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$ are independent and

$$\begin{aligned} &(n-p)\widehat{\sigma}^2\sigma^{-2} \sim \mathcal{X}^2(n-p), \\ &E(\widehat{\sigma}^2) = \sigma^2, \qquad \mathrm{var}(\widehat{\sigma}^2) = 2(n-p)^{-1}\sigma^4. \end{aligned} \tag{2.4}$$

Suppose that we have M groups of regressors. We combine M nested sub-models of (2.2) candidate models, and the mth candidate model includes the first m groups of variables of $\mathbf{X}$, denoted by $\mathbf{X}_m$. Denote the group size of the mth group by k_m. Let $v_m = \sum_{j=1}^{m} k_j$, and thus v_m is the number of variables used in the mth candidate model and is also the number of columns of $\mathbf{X}_m$.

Let $\boldsymbol{\Pi}_m$ be a selection matrix so that $\boldsymbol{\Pi}_m = (\mathbf{I}_{v_m}, \mathbf{0}_{v_m \times (p-v_m)})$ and thus $\mathbf{X}_m = \mathbf{X}\boldsymbol{\Pi}_m^{\mathrm{T}}$. Define a $p \times p$ matrix $\mathbf{A}_m = \boldsymbol{\Pi}_m^{\mathrm{T}}(\mathbf{X}_m{}^{\mathrm{T}}\mathbf{X}_m)^{-1}\boldsymbol{\Pi}_m(\mathbf{X}^{\mathrm{T}}\mathbf{X})$. Under the mth candidate model, the restricted OLS estimator of $\boldsymbol{\beta}$ can be written as

$$\widehat{\boldsymbol{\beta}}_m = \mathbf{A}_m\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}. \tag{2.5}$$

For the Mth candidate model, $\widehat{\boldsymbol{\beta}}_M = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. The model averaging estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m\widehat{\boldsymbol{\beta}}_m,$$

where w_m is the weight corresponding to the mth candidate model and $\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$, belonging to weight set $\mathcal{H} = \left\{\mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\right\}$.

Let $\mathbf{v} = (v_1, \ldots, v_M)^{\mathrm{T}}$. Hansen (2007) proposed choosing weights by minimizing Mallows' criterion

$$C(\mathbf{w}) = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}) - \mathbf{y}\right\|^2 + 2\widehat{\sigma}^2\mathbf{w}^{\mathrm{T}}\mathbf{v}. \tag{2.6}$$

Let $\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_M)^{\mathrm{T}} = \arg\min_{\mathbf{w}\in\mathcal{H}} C(\mathbf{w})$, so that the combined estimator $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ is the Mallows model averaging (MMA) estimator of $\boldsymbol{\beta}$.

3. MSE comparison

Let

$$\begin{aligned} q(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2) &= I(m_J < M)\left\{2\sigma^2(v_M - v_{m_J}) - \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2\right\} \\ &\quad + \sigma^4\sum_{j=1}^{J-1}\frac{\left\{(n-p-2)(n-p)^{-1}(v_{m_{j+1}} - v_{m_j}) - 4\right\}(v_{m_{j+1}} - v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}, \end{aligned}$$

where $I(\cdot)$ denotes the indicator function as usual. For any estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, its MSE is defined by $E\left\|\mathbf{X}\widetilde{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\right\|^2$.

Theorem 1. $E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \mathbf{X}\boldsymbol{\beta}\right\|^2 = \sigma^2 p - E\left\{q(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)\right\}$.

See Appendix A.1 for the proof of Theorem 1. From Theorem 1, we have the following result.

Corollary 1. *If* $(n-p-2)(n-p)^{-1}k_m \geq 4$ *for all* $m \geq 2$, *then*

$$E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \mathbf{X}\boldsymbol{\beta}\right\|^2 < E\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta}\right\|^2,$$

i.e., $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ *dominates* $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$.

See Appendix A.2 for the proof of Corollary 1. We note that the result in Corollary 1 provides the exact dominance condition for the MMA estimator over the OLS estimator in the MSE sense. This result is valid for any sample size n. In a special case, when n tending to infinity, Corollary 1 provides a sufficient condition for the MMA estimator dominating (asymptotically) the OLS estimator, which is $k_m \geq 4$ for all $m \geq 2$ (also see Hansen, 2014). Irrespective of sample size, our Corollary 1 indicates that there is a scale $(n-p-2)(n-p)^{-1} < 1$ associated with k_m and when $(n-p-2)(n-p)^{-1}k_m \geq 4$ for all $m \geq 2$, the MMA estimator dominates the OLS estimator.

4. Concluding remarks

Firstly we have developed new results on deriving the condition under which the MMA estimator dominates the OLS estimator in the exact MSE sense. This exact condition depends on the sample size and the number of regressors. Secondly we remark that our theory is also confined to the context of nested models. Extension of the current analysis to non-nested models will be very challenging. Lastly, the MSE comparison of the current paper is built under the normally distributed and homoscedastic error. Developing MSE comparison under other error cases is also an interesting topic for future research.

Acknowledgments

Zhang's research was supported by National Natural Science Foundation of China (Grant Nos. 71522004 and 11471324) and a special talent grant from AMSS, CAS. Also, Ullah's research was supported by the Academic Senate Grant at UCR. The work of Zhang was performed during his visit to Texas A&M University.

Appendix

A.1. Proof of Theorem 1

First, similar to Hansen (2014), we define cumulative weights

$$w_m^* = w_1 + \cdots + w_m$$

and $\mathbf{w}^* = (w_1^*, \ldots, w_M^*)^{\mathrm{T}}$. Then, $w_m = w_m^* - w_{m-1}^*$ for $m \geq 2$, $w_1 = w_1^*$, and $\mathbf{w} \in \mathcal{H}$ is equivalent to

$$\mathbf{w}^* \in \mathcal{H}^* \equiv \left\{\mathbf{w}^* \in [0,1]^M : 0 \leq w_1^* \leq \cdots \leq w_M^* = 1\right\} \tag{A.1}$$

and the model averaging estimator $\widehat{\boldsymbol{\beta}}(\mathbf{w})$ can be rewritten as

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \sum_{m=1}^{M-1} w_m^*(\widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m). \tag{A.2}$$

Let $\widehat{w}_m^* = \widehat{w}_1 + \cdots + \widehat{w}_m$, $\widehat{\mathbf{w}}^* = (\widehat{w}_1^*, \ldots, \widehat{w}_M^*)^{\mathrm{T}}$,

$$b_m = \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_m\right\|^2, \tag{A.3}$$

and

$$C^*(\mathbf{w}^*) = \sum_{m=1}^{M-1}\left\{w_m^{*2}(b_m - b_{m+1}) - 2\widehat{\sigma}^2 w_m^* k_{m+1}\right\}. \tag{A.4}$$

From Lemma 1 of Hansen (2014), we have

$$C(\mathbf{w}) = C^*(\mathbf{w}^*) + b_M + 2\widehat{\sigma}^2 v_M \tag{A.5}$$

and

$$\widehat{\mathbf{w}}^* = \arg\min_{\mathbf{w}^* \in \mathcal{H}^*} C^*(\mathbf{w}^*). \tag{A.6}$$

Hence, from (2.5) and (A.3)–(A.6), we know that both $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{w}}^*$ depend on $\mathbf{y}$ through $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$.

Since weights $\widehat{w}_1, \ldots, \widehat{w}_M$ are determined by data, the indexes of candidate models with positive weights are random. We use $\{m_1(\mathbf{y}), \ldots, m_{J(\mathbf{y})}(\mathbf{y})\}$ to denote the indexes set, where $J(\mathbf{y})$ and $m_{j(\mathbf{y})}(\mathbf{y})$ depend on $\mathbf{y}$. By the analysis of the above paragraph, we know that $C(\mathbf{w})$ depends on $\mathbf{y}$ through $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$, so we can write $J(\mathbf{y})$ and $m_{j(\mathbf{y})}(\mathbf{y})$ as $J(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)$ and $m_{j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)$. Instead, for simplicity, we write them as J and m_j, although they are random. Thus, $w^*_{m_j} = \cdots = w^*_{m_{j+1}-1}$ and $w^*_{m_J} = 1$. From the proof of Theorem 1 of Hansen (2014), we have

$$\begin{aligned}
C^*(\mathbf{w}^*) &= \sum_{m=1}^{M-1} \left\{ {w^*_m}^2 (b_m - b_{m+1}) - 2\widehat{\sigma}^2 w^*_m k_{m+1} \right\} \\
&= \sum_{j=1}^{J-1} \sum_{\ell=m_j}^{m_{j+1}-1} \left\{ w^{*2}_\ell (b_{\ell+1} - b_\ell) - 2\widehat{\sigma}^2 w^*_\ell k_{\ell+1} \right\} \\
&\quad + \sum_{\ell=m_J}^{M-1} \left\{ w^{*2}_\ell (b_{\ell+1} - b_\ell) - 2\widehat{\sigma}^2 w^*_\ell k_{\ell+1} \right\} \\
&= \sum_{j=1}^{J-1} \left\{ {w^*_{m_j}}^2 (b_{m_{j+1}} - b_{m_j}) - 2\widehat{\sigma}^2 w^*_{m_j} (v_{m_{j+1}} - v_{m_j}) \right\} \\
&\quad + (b_M - b_{m_J}) - 2\widehat{\sigma}^2 (v_M - v_{m_J}),
\end{aligned}$$

which is minimized by

$$\widehat{w}^*_{m_j} = \frac{\widehat{\sigma}^2 (v_{m_{j+1}} - v_{m_j})}{b_{m_{j+1}} - b_{m_j}}, \quad j = 1, \ldots, J-1, \tag{A.7}$$

when $w^* \in \mathcal{H}^*$ (see (A.1) for the definition of $\mathcal{H}^*$).

Note that the set $\{m_1, \ldots, m_J\}$ which contains indexes of candidate models with positive weights is random and depends on $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$. When $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$ vary, the set $\{m_1, \ldots, m_J\}$ can also vary, but it is a piecewise constant function of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$ and is almost differentiable in the sense of Stein (1981) except for a finite number of points. Hence, in the following proof, when taking derivatives with respect to $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$, we take $\{m_1, \ldots, m_J\}$ be a constant set.

Since model m_j is nested within model m_{j+1}, it is easily to obtain the following results

$$(\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})^{\mathrm{T}} (\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) = 0 \tag{A.8}$$

and

$$b_{m_{j+1}} - b_{m_j} = \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} \right\|^2 - \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2 = \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2. \tag{A.9}$$

It follows from (2.3), (2.5)–(A.7), (A.8)–(A.9), Stein Lemma (Stein, 1981), and the independence between $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$ that

$$\begin{aligned}
&E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}) - \mathbf{X}\boldsymbol{\beta} \right\|^2 \\
&= E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{m=1}^{M-1} \widehat{w}^*_m (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m+1} - \mathbf{X}\widehat{\boldsymbol{\beta}}_m) \right\|^2 \\
&= E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1} \sum_{\ell=m_j}^{m_{j+1}-1} \widehat{w}^*_\ell (\mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell+1} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\ell) \right. \\
&\quad \left. - I(m_J < M) \sum_{\ell=m_J}^{M-1} (\mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell+1} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\ell) \right\|^2 \\
&= E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1} \widehat{w}^*_{m_j} (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}) \right. \\
&\quad \left. - I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) \right\|^2 \\
&= E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1} \frac{\widehat{\sigma}^2 (v_{m_{j+1}} - v_{m_j})}{b_{m_{j+1}} - b_{m_j}} \right. \\
&\quad \left. \times (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}) - I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) \right\|^2 \\
&= E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta} - \sum_{j=1}^{J-1} \frac{\widehat{\sigma}^2 (v_{m_{j+1}} - v_{m_j})}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2} \right. \\
&\quad \left. \times (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}) - I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) \right\|^2 \\
&= E \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + E \sum_{j=1}^{J-1} \frac{\widehat{\sigma}^4 (v_{m_{j+1}} - v_{m_j})^2}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2} \\
&\quad + E \left\{ I(m_J < M) \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J} \right\|^2 \right\} \\
&\quad - 2E \left[\left\{ I(m_J < M)(\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) \right\}^{\mathrm{T}} \mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) \right] \\
&\quad - 2E \left[\left\{ \sum_{j=1}^{J-1} \frac{\widehat{\sigma}^2 (v_{m_{j+1}} - v_{m_j})}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2} (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}) \right\}^{\mathrm{T}} \right. \\
&\quad \left. \times \mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) \right] \\
&\quad + 2 \sum_{j=1}^{J-1} \frac{\widehat{\sigma}^2 (v_{m_{j+1}} - v_{m_j})}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2} (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})^{\mathrm{T}} \\
&\quad \times (\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}) I(m_J < M) \\
&= \sigma^2 p + E \left\{ \widehat{\sigma}^4 \sum_{j=1}^{J-1} \frac{(v_{m_{j+1}} - v_{m_j})^2}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2} \right\} \\
&\quad + E \left\{ I(m_J < M) \left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J} \right\|^2 \right\} \\
&\quad - 2\sigma^2 E \left[I(m_J < M) \mathrm{trace} \left\{ (\mathbf{A}_M - \mathbf{A}_{m_J})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \right\} \right] \\
&\quad - 2E \left(\widehat{\sigma}^2 E \left[\left\{ \sum_{j=1}^{J-1} \frac{(v_{m_{j+1}} - v_{m_j})}{\left\| \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j} \right\|^2} \right. \right. \right. \\
&\quad \left. \left. \left. \times (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}) \right\}^{\mathrm{T}} \mathbf{X}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} - \boldsymbol{\beta}) | \widehat{\sigma}^2 \right] \right) + 0
\end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 p + E\left\{\widehat{\sigma}^4 \sum_{j=1}^{J-1} \frac{(v_{m_{j+1}} - v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\} \\
&\quad + E\left[I(m_J < M)\left\{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2 - 2\sigma^2(v_M - v_{m_J})\right\}\right] \\
&\quad - 2\sigma^2 E\left[\widehat{\sigma}^2 \sum_{j=1}^{J-1} \frac{(v_{m_{j+1}} - v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right. \\
&\quad \left. \times \ \mathrm{trace}\left\{(\mathbf{A}_{m_{j+1}} - \mathbf{A}_{m_j})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\}\right] \\
&\quad + 4\sigma^2 E \sum_{j=1}^{J-1}\left\{\widehat{\sigma}^2 \frac{(v_{m_{j+1}} - v_{m_j})}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^4}\right\} \\
&\quad \times \mathrm{trace}\left\{(\mathbf{A}_{m_{j+1}}{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}_{m_{j+1}} - \mathbf{A}_{m_j}{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}_{m_j})\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right. \\
&\quad \left. \times \ (\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j})^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right\} \\
&= \sigma^2 p + E\left\{\widehat{\sigma}^4 \sum_{j=1}^{J-1} \frac{(v_{m_{j+1}} - v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\} \\
&\quad + E\left[I(m_J < M)\left\{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2 - 2\sigma^2(v_M - v_{m_J})\right\}\right] \\
&\quad - 2\sigma^2 E\left\{\widehat{\sigma}^2 \sum_{j=1}^{J-1} \frac{(v_{m_{j+1}} - v_{m_j})^2}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}\right\} \\
&\quad + 4\sigma^2 E \sum_{j=1}^{J-1} \widehat{\sigma}^2 \frac{v_{m_{j+1}} - v_{m_j}}{\left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_{m_{j+1}} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_j}\right\|^2}, \qquad \text{(A.10)}
\end{aligned}
$$

where the seventh and eighth "=" use Stein lemma.

Let $\widehat{a} = 2^{-1}(n-p)\widehat{\sigma}^2\sigma^{-2}$ and $a = 2^{-1}(n-p)$. From (2.4), we have

$$\widehat{a} \sim \mathrm{Gamma}\left\{2^{-1}(n-p), 1\right\} \qquad \text{(A.11)}$$

with mean a. So, by the independence between $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\sigma}^2$ and Lemma 2 of Shen and Huang (2006), we obtain that for any constant c and any functions of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, f_1(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}), \ldots, f_{J-1}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$,

$$
\begin{aligned}
&E\left\{\widehat{\sigma}^2 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} \\
&= E\left[E\left\{\widehat{\sigma}^2 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}\right] \\
&= \sigma^2 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} \\
&\quad + E\left[E\left\{(\widehat{\sigma}^2 - \sigma^2)\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}\right] \\
&= \sigma^2 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\}, \qquad \text{(A.12)}
\end{aligned}
$$

where the third "=" uses Lemma 2 of Shen and Huang (2006), and

$$
\begin{aligned}
&E\left\{\widehat{\sigma}^4 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} \\
&= E\left[E\left\{\widehat{\sigma}^4 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}\right] \\
&= \sigma^4 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} + 4\sigma^4(n-p)^{-2}E \\
&\quad \times \left[E\left\{(\widehat{a} - a)(\widehat{a} + a)\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}\right] \\
&= \sigma^4 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} + 4\sigma^4(n-p)^{-2} \\
&\quad \times E\left[E\left\{\widehat{a}\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}\right] \\
&= \sigma^4 E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\} + 2\sigma^2(n-p)^{-1}E \\
&\quad \times \left[E\left\{\widehat{\sigma}^2 \sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})|\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right\}\right] \\
&= \sigma^4\left\{1 + 2(n-p)^{-1}\right\} E\left\{\sum_{j=1}^{J-1}(v_{m_{j+1}} - v_{m_j})^c f_j(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})\right\}, \qquad \text{(A.13)}
\end{aligned}
$$

where the third "=" uses Lemma 2 of Shen and Huang (2006) and the fourth "=" uses (A.12). In addition,

$$
\begin{aligned}
&2(v_{m_{j+1}} - v_{m_j})^2 - 4(v_{m_{j+1}} - v_{m_j}) \\
&\quad - \left\{1 + 2(n-p)^{-1}\right\}(v_{m_{j+1}} - v_{m_j})^2 \\
&= (v_{m_{j+1}} - v_{m_j})\left\{2(v_{m_{j+1}} - v_{m_j})\right. \\
&\quad \left. - 4 - (n-p+2)(n-p)^{-1}(v_{m_{j+1}} - v_{m_j})\right\} \\
&= (v_{m_{j+1}} - v_{m_j})\left\{(n-p-2)(n-p)^{-1}(v_{m_{j+1}} - v_{m_j}) - 4\right\}. \qquad \text{(A.14)}
\end{aligned}
$$

The result of Theorem 1 is implied by the above (A.10)–(A.14) and the definition of $q(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)$.

A.2. Proof of Corollary 1

From (A.11) and Lemma 2 of Shen and Huang (2006), we have

$$E\left\{(\widehat{\sigma}^2 - \sigma^2)I(m_J < M)(v_M - v_{m_J})\right\} = 0. \qquad \text{(A.15)}$$

Similar to the derivation of (28) in Hansen (2014), we can obtain that if $m_J < M$, then

$$b_M - b_{m_J} \le \widehat{\sigma}^2(v_M - v_{m_J}). \qquad \text{(A.16)}$$

It is seen from (A.9), (A.15)–(A.16) and the definition of $q(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)$ that when $(n-p-2)(n-p)^{-1}k_m \ge 4$ for all $m \ge 2$, we have

$$
\begin{aligned}
&E\left\{q(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \widehat{\sigma}^2)\right\} \\
&\ge E\left[I(m_J < M)\left\{2\sigma^2(v_M - v_{m_J}) - \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}_M - \mathbf{X}\widehat{\boldsymbol{\beta}}_{m_J}\right\|^2\right\}\right] \\
&= E\left[I(m_J < M)\left\{\widehat{\sigma}^2(v_M - v_{m_J}) - b_M - b_{m_J}\right\}\right] \\
&\quad - E\left\{I(m_J < M)(\widehat{\sigma}^2 - \sigma^2)(v_M - v_{m_J})\right\} \\
&\quad + \sigma^2 E\left\{I(m_J < M)(v_M - v_{m_J})\right\} \\
&\ge \sigma^2 E\left\{I(m_J < M)(v_M - v_{m_J})\right\}. \qquad \text{(A.17)}
\end{aligned}
$$

When $m_J < M$, $I(m_J < M)(v_M - v_{m_J})$ is larger than zero, which, along with (A.17), implies the result of Corollary 1.

References

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

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

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

Hansen, B.E., 2010. Averaging estimators for autoregressions with a near unit root. J. Econometrics 158, 142–155.

Hansen, B.E., 2014. Model averaging, asymptotic risk, and regressor groups. Quant. Econom. 5, 495–530.

Hansen, B.E., Racine, J., 2012. Jacknife model averaging. J. Econometrics 167, 38–46.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Amer. Statist. Assoc. 98, 879–899.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Statist. Sci. 14, 382–417.

James, W., Stein, C., 1961. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 361–380.

Kuersteiner, G., Okui, R., 2010. Constructing optimal instruments by first-stage prediction averaging. Econometrica 78, 697–718.

Li, D., Linton, O., Lu, Z., 2015. A flexible semiparametric forecasting model for time series. J. Econometrics 187, 345–357.

Liang, H., Zou, G., Wan, A.T.K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. J. Amer. Statist. Assoc. 106, 1053–1066.

Liu, C.-A., 2015. Distribution theory of the least squares averaging estimator. J. Econometrics 186, 142–159.

Liu, Q., Okui, R., 2013. Heteroskedasticity-robust Cp model averaging. Econom. J. 16, 462–473.

Lu, X., Su, L., 2015. Jackknife model averaging for quantile regressions. J. Econometrics 188, 40–58.

Shen, X., Huang, H.-C., 2006. Optimal model assessment, selection, and combination. J. Amer. Statist. Assoc. 101, 554–568.

Stein, C., 1981. Estimation of the mean of a multivariate normal distribution. Ann. Statist. 153, 1135–1151.

Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Ann. Statist. 39, 174–200.

Zhang, X., Wan, A., Zhou, S.Z., 2012. Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. J. Bus. Econom. Statist. 30, 132–142.

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