MPRA
Munich Personal RePEc Archive

A plug-in averaging estimator for regressions with heteroskedastic errors

CHU-AN LIU

National University of Singapore

10. August 2012

Online at http://mpra.ub.uni-muenchen.de/41414/
MPRA Paper No. 41414, posted 18. September 2012 13:52 UTC

A Plug-In Averaging Estimator for Regressions with Heteroskedastic Errors

Chu-An Liu*
National University of Singapore†

ecslca@nus.edu.sg
August 10, 2012

Abstract

This paper proposes a new model averaging estimator for the linear regression model with heteroskedastic errors. We address the issues of how to optimally assign the weights for candidate models and how to make inference based on the averaging estimator. We derive the asymptotic mean squared error (AMSE) of the averaging estimator in a local asymptotic framework, and then choose the optimal weights by minimizing the AMSE. We propose a plug-in estimator of the optimal weights and use these estimated weights to construct a plug-in averaging estimator of the parameter of interest. We derive the asymptotic distribution of the plug-in averaging estimator and suggest a plug-in method to construct confidence intervals. Monte Carlo simulations show that the plug-in averaging estimator has much smaller expected squared error, maximum risk, and maximum regret than other existing model selection and model averaging methods. As an empirical illustration, the proposed methodology is applied to cross-country growth regressions.

Keywords: Local asymptotic theory, Model averaging, Model selection, Plug-in estimators.
JEL Classification: C51, C52.

*I am deeply indebted to Bruce Hansen and Jack Porter for guidance and encouragement. I also thank Xiaoxia Shi, Biing-Shen Kuo, Yu-Chin Hsu for helpful discussions.

†Department of Economics, National University of Singapore, AS2 Level 6, 1 Arts Link, 117570 Singapore.

1 Introduction

In recent years, interest has increased in model averaging from the frequentist perspective. Unlike model selection, which picks a single model among the candidate models, model averaging incorporates all available information by averaging over all potential models. Model averaging is more robust than model selection since the averaging estimator considers the uncertainty across different models as well as the model bias from each candidate model. The central questions of concern are how to optimally assign the weights for candidate models and how to make inference based on the averaging estimator. This paper proposes a plug-in averaging estimator to resolve both of these issues. We derive the asymptotic mean squared error (AMSE) of the averaging estimator in a local asymptotic framework. We show that the optimal model weights which minimize the AMSE depend on the local parameters and the covariance matrix. The idea of the plug-in averaging estimator is to estimate the infeasible optimal weights by minimizing the sample analog of the AMSE. We show that the plug-in averaging estimator has a non-standard asymptotic distribution. Hence, confidence intervals based on normal approximations lead to distorted inference in this context. We suggest a plug-in method to construct confidence intervals, which have good finite-sample coverage probabilities.

Empirical studies often must consider whether additional regressors should be included in the baseline model. Adding more regressors reduces the model bias but causes a large variance. To address the trade-off between bias and variance, this paper studies model averaging in a local asymptotic framework where the regression coefficients are in a local $n^{-1/2}$ neighborhood of zero. Under drifting sequences of parameters, the AMSE of the averaging estimator remains finite and provides a good approximation to the finite sample MSE. The $O(n^{-1/2})$ framework is canonical in the sense that both squared model biases and estimator variances have the same order $O(n^{-1})$. Therefore, the optimal model is the one that has the best trade-off between squared model biases and estimator variances. The local-to-zero framework is crucial to analyze the asymptotic distribution of the averaging estimator. If all regression coefficients are fixed, then the model bias term tends to infinity and dominates the limiting distribution. In such a situation, the model which includes all regressors is the only one we should consider. The local asymptotic framework also implies that all of the candidate models are close to each other as the sample size increases. Hence, it is informative to employ model averaging rather than model selection in this framework.

We first consider the fixed weights for candidate models and then derive the asymptotic distribution of the averaging estimator in a local asymptotic framework, which allows us to characterize the optimal weights. The optimal weights are found by numerical minimization of the AMSE. We propose a plug-in estimator of the infeasible optimal weights. The optimal weights cannot be estimated consistently because they depend on the local parameters which cannot be estimated consistently. Estimated weights are asymptotically random, and this must be taken into account in the asymptotic distribution of the plug-in averaging estimator. To address this issue, we first show the joint convergence in distribution of all candidate models and the data-driven weights. Then, we derive the asymptotic distribution of the plug-in estimator, which is a non-linear function of the

normal random vector.

In addition to the plug-in averaging estimator, we also derive the asymptotic distributions of the Akaike information criterion (AIC) selection estimator (Akaike, 1973), the smoothed AIC (S-AIC) model averaging estimator (Buckland, Burnham, and Augustin, 1997), and the Jackknife Model Averaging (JMA) estimator (Hansen and Racine, 2012) in the local asymptotic framework. Although the asymptotic distribution of the averaging estimator with data-driven weights is non-standard, it can be approximated by simulation. Numerical comparisons show that the plug-in averaging estimator has substantially smaller risk than other data-driven averaging estimators in most ranges of the parameter space.

The empirical literature tends to focus on one particular parameter instead of assessing the overall properties of the model. In contrast to most existing model selection and model averaging methods, our method is tailored to the parameter of interest. The proposed averaging estimator is constructed based on the focus parameter instead of the global fit of the model. The focus parameter is a smooth real-valued function of regression coefficients. Thus, we focus attention on a low-dimension function of the model parameters. Also, we allow different model weights to be chosen for different parameters of interest.

One straightforward way to construct the confidence interval for the focus parameter is to employ the t-statistic. The confidence interval is constructed by inverting the t-statistic based on the parameter of interest. We show that the asymptotic distribution of the model averaging t-statistic depends on unknown local parameters, and thus cannot be directly used for inference. We propose a plug-in method to construct the confidence interval based on a non-standard limiting distribution. The idea is to simulate the limiting distribution of the model averaging t-statistic by replacing the unknown parameters with plug-in estimators. The confidence interval is constructed based on the $1-\alpha$ quantile of the simulated distribution. Our simulations show that the coverage probability of the plug-in confidence interval is close to the nominal level, while the confidence interval based on normal approximations leads to distorted inference.

There is a growing body of literature on frequentist model averaging. Buckland, Burnham, and Augustin (1997) suggest selecting the weights using the exponential AIC. Yang (2001) and Yuan and Yang (2005) propose an adaptive regression by mixing models. Hansen (2007) introduces the Mallows Model Averaging estimator for nested and homoskedastic models where the weights are selected by minimizing the Mallows criterion. Wan, Zhang, and Zou (2010) extend the asymptotical optimality of the Mallows Model Averaging estimator for continuous weights and a non-nested set-up. Hansen and Racine (2012) propose the Jackknife Model Averaging estimator for non-nested and heteroskedastic models where the weights are chosen by minimizing a leave-one-out cross-validation criterion. Liang, Zou, Wan, and Zhang (2011) suggest selecting the weights by minimizing the trace of an unbiased estimator of mean squared error. These papers propose methods of determining weights without deriving the asymptotic distribution of the proposed estimator, which is difficult to make inference based on their estimators. In contrast to frequentist model averaging, there is a large body of literature on Bayesian model averaging (see Hoeting, Madigan, Raftery, and Volinsky (1999) for a literature review).

The idea of using the local asymptotic framework to investigate the limiting distributions of model averaging estimators is developed by Hjort and Claeskens (2003) and Claeskens and Hjort (2008). However, their work is limited to the likelihood-based model. Following Hjort and Claeskens (2003), DiTraglia (2011) proposes a moment selection criterion and a moment averaging estimator for the GMM framework. Like DiTraglia, we employ a drifting asymptotic framework to approximate the finite sample MSE. Unlike DiTraglia, we consider model averaging rather than moment averaging, and we combine the models with valid moment conditions rather than potentially invalid moment conditions. Other work on the asymptotic properties of averaging estimators includes Leung and Barron (2006), Pötscher (2006), and Hansen (2009, 2010). Leung and Barron (2006) study the risk bound of the averaging estimator under a normal error assumption. Pötscher (2006) analyzes the finite sample and asymptotic distributions of the averaging estimator for the two-model case. Hansen (2009) evaluates the AMSE of averaging estimators for the linear regression model with a possible structural break. Hansen (2010) examines the AMSE and forecast expected squared error of averaging estimators in an autoregressive model with a near unit root in a local-to-unity framework. Most of these studies, however, are limited to the two-model case and the homoskedastic framework.

There is a large literature on inference after model selection, including Pötscher (1991), Kabaila (1995, 1998), Leeb and Pötscher (2003, 2005, 2006, 2008). These papers point out that the coverage probability of the confidence interval based on the model selection estimator is lower than the nominal level. They also argue that the conditional and unconditional distribution of post-model-selection estimators cannot be uniformly consistently estimated. In the model averaging literature, Hjort and Claeskens (2003) and Claeskens and Hjort (2008) show that the traditional confidence interval based on normal approximations leads to distorted inference. Pötscher (2006) argues that the finite-sample distribution of the averaging estimator cannot be uniformly consistently estimated.

There are also alternatives to model selection and model averaging. Tibshirani (1996) introduces the LASSO estimator, a method for simultaneous estimation and variable selection. Zou (2006) proposes the adaptive LASSO approach and presents its oracle properties. White and Lu (2010) propose a new Hausman (1978) type test of robustness for the core regression coefficients. They also provide a feasible optimally combined GLS estimator. Hansen, Lunde, and Nason (2011) propose the model confidence set which is constructed based on an equivalence test.

The outline of the paper is as follows. Section 2 presents the model and the averaging estimator of the focus parameter. Section 3 presents the asymptotic distribution of the averaging estimator with fixed weights in a local asymptotic framework. Section 4 introduces the plug-in averaging estimator and derives the limiting distribution. Section 5 presents the asymptotic distributions of AIC, S-AIC and JMA estimators. The results of the two-model case are presented. Section 6 evaluates the finite sample properties of the plug-in averaging estimator and other averaging estimators. Section 7 discusses the confidence interval construction. Section 8 applies the plug-in averaging estimator to cross-country growth regressions. Section 9 concludes. Proofs, figures, and tables are included in the Appendix.

2 Model and Estimation

Consider a linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + e_i, \tag{2.1}$$

$$\mathrm{E}(e_i|\mathbf{x}_i, \mathbf{z}_i) = 0, \tag{2.2}$$

$$\mathrm{E}(e_i^2|\mathbf{x}_i, \mathbf{z}_i) = \sigma^2(\mathbf{x}_i, \mathbf{z}_i), \tag{2.3}$$

where y_i is a scalar dependent variable, $\mathbf{x}_i = (x_{1i}, ..., x_{ki})'$ and $\mathbf{z}_i = (z_{1i}, ..., z_{\ell i})'$ are vectors of regressors, e_i is an unobservable regression error, and $\boldsymbol{\beta}(k \times 1)$ and $\boldsymbol{\gamma}(\ell \times 1)$ are unknown parameter vectors. The error term is allowed to be heteroskedastic and there is no further assumption on the distribution of the error term. Here, $\mathbf{x}_i$ are the core regressors which must be included in the model based on theoretical grounds, while $\mathbf{z}_i$ are the auxiliary regressors which may or may not be included in the model. Note that $\mathbf{x}_i$ may only include a constant term or even an empty matrix. In matrix notation, we write the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e} \tag{2.4}$$

where $\mathbf{H} = (\mathbf{X}, \mathbf{Z})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$.

The parameter of interest is $\mu = \mu(\boldsymbol{\theta}) = \mu(\boldsymbol{\beta}, \boldsymbol{\gamma})$, which is a smooth real-valued function. Unlike the traditional model selection and model averaging approaches which assess the global fit of the model, we evaluate the model based on the focus parameter μ. For example, μ may be an individual coefficient or a ratio of two coefficients of regressors.

Let M be the number of submodels, where the submodel includes all core regressors $\mathbf{X}$ and a subset of auxiliary regressors $\mathbf{Z}$. The m'th submodel has $k + \ell_m$ regressors. If we consider a sequence of nested models, then $M = \ell + 1$. If we consider all possible submodels, then $M = 2^\ell$. Let $\boldsymbol{\Pi}_m$ be the $\ell_m \times \ell$ selection matrix which selects the included auxiliary regressors. Here, ℓ_m is the number of auxiliary regressors $\mathbf{z}_i$ included in the submodel m.

The least-squares estimator of $\boldsymbol{\theta}$ for the full model, i.e. all auxiliary regressors are included in the model, is

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y}, \tag{2.5}$$

and the estimator for the submodel m is

$$\tilde{\boldsymbol{\theta}}_m = \begin{pmatrix} \tilde{\boldsymbol{\beta}}_m \\ \tilde{\boldsymbol{\gamma}}_m \end{pmatrix} = (\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{y}, \tag{2.6}$$

where $\mathbf{H}_m = (\mathbf{X}, \mathbf{Z}\boldsymbol{\Pi}_m')$ with $m = 1, ..., M$. Let $\mathbf{I}$ denote an identity matrix and $\mathbf{0}$ a zero matrix. If $\boldsymbol{\Pi}_m = \mathbf{I}_\ell$, then we have $\tilde{\boldsymbol{\theta}}_m = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{y} = \hat{\boldsymbol{\theta}}$, the least-squares estimator for the full model. If $\boldsymbol{\Pi}_m = \mathbf{0}$, then we have $\tilde{\boldsymbol{\theta}}_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the least-squares estimator for the narrow model, or the smallest model among all possible submodels.

We now define the averaging estimator of the focus parameter μ. Let $\mathbf{w} = (w_1, ..., w_M)'$ be a weight vector with $w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$. That is, the weight vector lies in the unit simplex in $\mathbb{R}^M$:

$$\mathcal{H}_n = \left\{ \mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}.$$

The sum of the weight vector is required to be one. Otherwise, the averaging estimator is not consistent. Let $\tilde{\mu}_m = \mu(\tilde{\boldsymbol{\theta}}_m) = \mu(\tilde{\boldsymbol{\beta}}_m, \tilde{\boldsymbol{\gamma}}_m)$ denote the submodel estimates. The averaging estimator of μ is

$$\bar{\mu}(\mathbf{w}) = \sum_{m=1}^{M} w_m \tilde{\mu}_m. \tag{2.7}$$

Here we want to point out that we have less restrictions on the weight function than other existing methods. Leung and Barron (2006), Pötscher (2006), and Liang, Zou, Wan, and Zhang (2011) assume the parametric form of the weight function. Hansen (2007) and Hansen and Racine (2012) restrict the weights to be discrete. Contrary to these works, we allow continuous weights without assuming any parametric form, which is more general and applicable than other approaches.

3 Asymptotic Properties

To establish the asymptotic distribution of the averaging estimator, we follow Hjort and Claeskens (2003) and use a local-to-zero asymptotic framework where the auxiliary parameters $\boldsymbol{\gamma}$ are in a local $n^{-1/2}$ neighborhood of zero. Let $\mathbf{h}_i = (\mathbf{x}_i', \mathbf{z}_i')'$ and $\mathbf{Q} = \mathrm{E}\left(\mathbf{h}_i \mathbf{h}_i'\right)$ partitioned so that $\mathrm{E}\left(\mathbf{x}_i \mathbf{x}_i'\right) = \mathbf{Q}_{\mathbf{xx}}$, $\mathrm{E}\left(\mathbf{x}_i \mathbf{z}_i'\right) = \mathbf{Q}_{\mathbf{xz}}$, and $\mathrm{E}\left(\mathbf{z}_i \mathbf{z}_i'\right) = \mathbf{Q}_{\mathbf{zz}}$. Let $\boldsymbol{\Omega} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{E}\big(\mathbf{h}_i \mathbf{h}_j' e_i e_j\big)$ partitioned so that $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{E}\big(\mathbf{x}_i \mathbf{x}_j' e_i e_j\big) = \boldsymbol{\Omega}_{\mathbf{xx}}$, $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{E}\big(\mathbf{x}_i \mathbf{z}_j' e_i e_j\big) = \boldsymbol{\Omega}_{\mathbf{xz}}$, and $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{E}\big(\mathbf{z}_i \mathbf{z}_j' e_i e_j\big) = \boldsymbol{\Omega}_{\mathbf{zz}}$. Note that if the error term e_i is serially uncorrelated, $\boldsymbol{\Omega}$ can be simplified as $\boldsymbol{\Omega} = \mathrm{E}\left(\mathbf{h}_i \mathbf{h}_i' e_i^2\right)$.

Assumption 1. As $n \to \infty$, $n^{1/2}\boldsymbol{\gamma} = n^{1/2}\boldsymbol{\gamma}_n \to \boldsymbol{\delta} \in \mathbb{R}^{\ell}$.

Assumption 2. As $n \to \infty$, $n^{-1}\mathbf{H}'\mathbf{H} \xrightarrow{p} \mathbf{Q}$ and $n^{-1/2}\mathbf{H}'\mathbf{e} \xrightarrow{d} \mathbf{R} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$.

Assumption 1 is the key assumption to develop the asymptotic distribution. It is a common assumption in the weak instrument literature, see Staiger and Stock (1997). This assumption says the partial correlations between the auxiliary regressors and the dependent variable are weak. This assumption implies that as the sample size increases, all of the submodels are close to each other. Under this framework, it is informative to know if we can do better by averaging the candidate models, instead of choosing one single model. Also note that the $O(n^{-1/2})$ framework gives squared model biases of the same order $O(n^{-1})$ as estimator variances. Hence, the optimal model is the one that achieve the best trade-off between bias and variance.

Assumption 2 is a high-level condition which permits the application of cross-section, panel, and time-series data. This condition holds under appropriate primitive assumptions. For example, if y_i is a stationary and ergodic martingale difference sequence with finite fourth moments, then the condition follows from the weak law of large numbers and the central limit theorem for martingale difference sequences.

Since the selection matrix $\mathbf{\Pi}_m$ is non-random with elements either 0 or 1, for the submodel m we have $n^{-1}\mathbf{H}_m'\mathbf{H}_m \xrightarrow{p} \mathbf{Q}_m$ where $\mathbf{Q}_m$ is nonsingular with

$$\mathbf{Q}_m = \begin{pmatrix} \mathbf{Q_{xx}} & \mathbf{Q_{xz}\Pi}_m' \\ \mathbf{\Pi}_m\mathbf{Q_{zx}} & \mathbf{\Pi}_m\mathbf{Q_{zz}\Pi}_m' \end{pmatrix},$$

and $n^{-1/2}\mathbf{H}_m'\mathbf{e} \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{\Omega}_m)$ with

$$\mathbf{\Omega}_m = \begin{pmatrix} \mathbf{\Omega_{xx}} & \mathbf{\Omega_{xz}\Pi}_m' \\ \mathbf{\Pi}_m\mathbf{\Omega_{zx}} & \mathbf{\Pi}_m\mathbf{\Omega_{zz}\Pi}_m' \end{pmatrix}.$$

Let $\boldsymbol{\theta}_m = (\boldsymbol{\beta}', \boldsymbol{\gamma}_m')' = (\boldsymbol{\beta}', \boldsymbol{\gamma}'\mathbf{\Pi}_m')'$. In this section, we concentrate on fixed weights. The averaging estimator with data-driven weights is presented in the next section. The following lemmas describe the asymptotic distributions of the least-squares estimators and the limiting distribution of the focus parameter.

Lemma 1. *Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathbf{Q}^{-1}\mathbf{R} \sim \mathbf{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\right),$$
$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m\right) \xrightarrow{d} \mathbf{A}_m\boldsymbol{\delta} + \mathbf{B}_m\mathbf{R} \sim \mathbf{N}\left(\mathbf{A}_m\boldsymbol{\delta},\ \mathbf{Q}_m^{-1}\mathbf{\Omega}_m\mathbf{Q}_m^{-1}\right),$$

where

$$\mathbf{A}_m = \mathbf{Q}_m^{-1}\begin{pmatrix} \mathbf{Q_{xz}} \\ \mathbf{\Pi}_m\mathbf{Q_{zz}} \end{pmatrix}\left(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m\right),\ \mathbf{B}_m = \mathbf{Q}_m^{-1}\mathbf{S}_m',\ \textit{and}\ \mathbf{S}_m = \begin{pmatrix} \mathbf{I}_k & \mathbf{0}_{k\times\ell_m} \\ \mathbf{0}_{\ell\times k} & \mathbf{\Pi}_m' \end{pmatrix}.$$

Note that $\mathbf{S}_m$ is an extended selection matrix of dimension $(k+\ell)\times(k+\ell_m)$. Denote $\mathbf{D}_{\boldsymbol{\theta}_m} = \left(\mathbf{D}_{\boldsymbol{\beta}}', \mathbf{D}_{\boldsymbol{\gamma}_m}'\right)'$, $\mathbf{D}_{\boldsymbol{\beta}} = \partial\mu/\partial\boldsymbol{\beta}$, and $\mathbf{D}_{\boldsymbol{\gamma}_m} = \partial\mu/\partial\boldsymbol{\gamma}_m$ with partial derivatives evaluated at the null points $(\boldsymbol{\beta}', \mathbf{0}')'$.

Lemma 2. *Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\sqrt{n}\left(\mu(\tilde{\boldsymbol{\theta}}_m) - \mu(\boldsymbol{\theta})\right) \xrightarrow{d} \Lambda_m = \mathbf{a}_m'\boldsymbol{\delta} + \mathbf{b}_m'\mathbf{R} \sim \mathbf{N}\left(\mathbf{a}_m'\boldsymbol{\delta},\ \mathbf{D}_{\boldsymbol{\theta}_m}'\mathbf{Q}_m^{-1}\mathbf{\Omega}_m\mathbf{Q}_m^{-1}\mathbf{D}_{\boldsymbol{\theta}_m}\right),$$

where

$$\mathbf{a}_m = \left(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m\right)\left(\begin{pmatrix} \mathbf{Q_{zx}} \\ \mathbf{Q_{zz}\Pi}_m' \end{pmatrix}\mathbf{Q}_m^{-1}\mathbf{D}_{\boldsymbol{\theta}_m} - \mathbf{D}_{\boldsymbol{\gamma}}\right)\ \textit{and}\ \mathbf{b}_m = \mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{D}_{\boldsymbol{\theta}_m}.$$

The main difference between Lemma 1 and 2 is the asymptotic distribution of the focus parameter involves the partial derivatives. Note that both $\mathbf{A}_m\boldsymbol{\delta}$ and $\mathbf{a}_m'\boldsymbol{\delta}$ represent the bias terms of submodel estimators. To be more precise, the biases come from the omitted auxiliary regressors. As we can see from $(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m)$, this is the selection matrix which selects the omitted auxiliary regressors.

Lemma 1 and 2 imply joint convergence in distribution of all submodels since all asymptotic distributions of submodels can be expressed in terms of the same normal random vector $\mathbf{R}$. The following theorem shows the asymptotic normality of the averaging estimator with fixed weights.

Theorem 1. *Suppose Assumptions 1-2 hold. As $n \to \infty$, we have*

$$\sqrt{n}\left(\bar{\mu}(\mathbf{w}) - \mu\right) \xrightarrow{d} \mathbf{N}\left(\mathbf{a}'\boldsymbol{\delta}, V\right)$$

where

$$\begin{aligned}
\mathbf{a} &= \sum_{m=1}^{M} w_m \mathbf{a}_m, \\
V &= \sum_{m=1}^{M} w_m^2 \mathbf{D}_{\boldsymbol{\theta}_m}' \mathbf{Q}_m^{-1} \mathbf{\Omega}_m \mathbf{Q}_m^{-1} \mathbf{D}_{\boldsymbol{\theta}_m} + 2\sum_{m<p} w_m w_p \mathbf{D}_{\boldsymbol{\theta}_m}' \mathbf{Q}_m^{-1} \mathbf{\Omega}_{m,p} \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p}, \\
\mathbf{\Omega}_{m,p} &= \begin{pmatrix} \mathbf{\Omega}_{\mathbf{xx}} & \mathbf{\Omega}_{\mathbf{xz}}\mathbf{\Pi}_p' \\ \mathbf{\Pi}_m\mathbf{\Omega}_{\mathbf{zx}} & \mathbf{\Pi}_m\mathbf{\Omega}_{\mathbf{zz}}\mathbf{\Pi}_p' \end{pmatrix},
\end{aligned}$$

and $\mathbf{a}_m$ is defined in Lemma 2.

Following by Theorem 1, we can derive the AMSE of the averaging estimator. Here, we define the AMSE as $\mathrm{AMSE}(\hat{\mu}) = \lim_{n\to\infty} \mathrm{E}\big(n(\hat{\mu}-\mu)^2\big)$. Then the AMSE of the averaging estimator (2.7) is

$$\mathrm{AMSE}(\bar{\mu}(\mathbf{w})) = \mathbf{w}'\boldsymbol{\zeta}\mathbf{w} \tag{3.1}$$

where $\boldsymbol{\zeta}$ is an $M \times M$ matrix with the (m,p)th element

$$\zeta_{m,p} = \boldsymbol{\delta}'\mathbf{a}_m\mathbf{a}_p'\boldsymbol{\delta} + \mathbf{D}_{\boldsymbol{\theta}_m}'\mathbf{Q}_m^{-1}\mathbf{\Omega}_{m,p}\mathbf{Q}_p^{-1}\mathbf{D}_{\boldsymbol{\theta}_p} \tag{3.2}$$

where $\mathbf{a}_m$ is defined in Lemma 2 and $\mathbf{\Omega}_{m,p}$ is defined in Theorem 1.

The optimal fixed-weight vector is the value which minimizes $\mathrm{AMSE}(\bar{\mu}(\mathbf{w}))$ over $\mathbf{w} \in \mathcal{H}_n$:

$$\mathbf{w}^o = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\zeta}\mathbf{w}. \tag{3.3}$$

Although there is no closed-form solution to (3.3) when $M > 2$, the weight vector can be found numerically via quadratic programming for which numerical algorithms are available for most programming languages. The minimized AMSE gives a benchmark to compare the AMSE and MSE of data-driven averaging estimators.

4 Plug-In Averaging Estimator

The optimal fixed weights derived in the previous section are infeasible, since they depend on the unknown parameters, $\mathbf{D}_{\boldsymbol{\theta}_m}$, $\mathbf{Q}_m$, $\boldsymbol{\Omega}_{m,p}$, $\mathbf{a}_m$, and $\boldsymbol{\delta}$. Furthermore, the optimal fixed weights cannot be estimated directly because there is no closed form expression when the number of models is greater than two. A straightforward solution is to estimate the AMSE of the averaging estimator given in (3.1) and (3.2), and to choose the data-driven weights by minimizing the sample analog of the AMSE.

The plug-in estimator of AMSE($\bar{\mu}(\mathbf{w})$) is $\mathbf{w}'\hat{\boldsymbol{\zeta}}\mathbf{w}$ where $\hat{\boldsymbol{\zeta}}$ is the sample analog of $\boldsymbol{\zeta}$ with the (m,p)th element

$$\hat{\zeta}_{m,p} = \hat{\boldsymbol{\delta}}'\hat{\mathbf{a}}_m\hat{\mathbf{a}}_p'\hat{\boldsymbol{\delta}} + \hat{\mathbf{D}}_{\boldsymbol{\theta}_m}'\hat{\mathbf{Q}}_m^{-1}\hat{\boldsymbol{\Omega}}_{m,p}\hat{\mathbf{Q}}_p^{-1}\hat{\mathbf{D}}_{\boldsymbol{\theta}_p}.$$

The weight vector of the plug-in estimator is defined as

$$\hat{\mathbf{w}}_{pia} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}'\hat{\boldsymbol{\zeta}}\mathbf{w}. \tag{4.1}$$

The plug-in averaging estimator is

$$\bar{\mu}(\hat{\mathbf{w}}_{pia}) = \sum_{m=1}^{M} \hat{w}_{pia,m}\tilde{\mu}_m. \tag{4.2}$$

We now discuss the plug-in estimator $\hat{\zeta}_{m,p}$. We first consider the estimator of $\mathbf{D}_{\boldsymbol{\theta}_m}$. Let $\hat{\mathbf{D}}_{\boldsymbol{\theta}_m} = \mathbf{S}_m'\hat{\mathbf{D}}_{\boldsymbol{\theta}}$ and $\hat{\mathbf{D}}_{\boldsymbol{\theta}} = \partial\mu(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ where $\mathbf{S}_m$ defined in Lemma 1 is a non-random selection matrix and $\hat{\boldsymbol{\theta}}$ is the estimate from the full model. By Lemma 1 and the continuous mapping theorem, it follows that $\hat{\mathbf{D}}_{\boldsymbol{\theta}_m}$ is a consistent estimator of $\mathbf{D}_{\boldsymbol{\theta}_m}$.

Next, we consider the estimators of $\mathbf{Q}_m$, $\boldsymbol{\Omega}_{m,p}$, and $\mathbf{a}_m$. Let $\hat{\mathbf{Q}}_m = \mathbf{S}_m'\hat{\mathbf{Q}}\mathbf{S}_m$, $\hat{\boldsymbol{\Omega}}_{m,p} = \mathbf{S}_m'\hat{\boldsymbol{\Omega}}\mathbf{S}_p$, and

$$\hat{\mathbf{a}}_m = \left(\mathbf{I}_\ell - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\left(\begin{pmatrix} \hat{\mathbf{Q}}_{\mathbf{zx}} \\ \hat{\mathbf{Q}}_{\mathbf{zz}}\boldsymbol{\Pi}_m' \end{pmatrix}\hat{\mathbf{Q}}_m^{-1}\hat{\mathbf{D}}_{\boldsymbol{\theta}_m} - \hat{\mathbf{D}}_{\boldsymbol{\gamma}}\right). \tag{4.3}$$

Consistent estimators for $\mathbf{Q}_m$, $\boldsymbol{\Omega}_{m,p}$, and $\mathbf{a}_m$ are available, since these unknown parameters are functions of the covariance matrix $\mathbf{Q}$ and $\boldsymbol{\Omega}$. We use the method of moments estimators for $\mathbf{Q}$ and $\boldsymbol{\Omega}$. Let $\hat{\mathbf{Q}} = n^{-1}\sum_{i=1}^n \mathbf{h}_i\mathbf{h}_i'$ and it follows that $\hat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$. If the error term e_i is serially uncorrelated, then $\boldsymbol{\Omega}$ can be estimated consistently by the heteroskedasticity-consistent covariance matrix estimator

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_i\mathbf{h}_i'\hat{e}_i^2, \tag{4.4}$$

which is proposed by White (1980). Here $\hat{e}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \mathbf{z}_i'\hat{\boldsymbol{\gamma}}$ is the least squares residual from the full model. If the error term e_i is serially correlated, then $\boldsymbol{\Omega}$ can be estimated consistently by the

heteroskedasticity and autocorrelation consistent covariance matrix estimator

$$\hat{\mathbf{\Omega}} = \sum_{j=-n}^{n} k(j/S_n)\hat{\mathbf{\Gamma}}(j), \tag{4.5}$$

$$\hat{\mathbf{\Gamma}}(j) = \frac{1}{n}\sum_{i=1}^{n-j} \mathbf{h}_i\mathbf{h}'_{i+j}\hat{e}_i\hat{e}_{i+j}, \text{ for } j \geq 0, \tag{4.6}$$

$$\hat{\mathbf{\Gamma}}(j) = \hat{\mathbf{\Gamma}}(-j)', \text{ for } j < 0, \tag{4.7}$$

where $k(\cdot)$ is a kernel function and S_n the bandwidth. Under some regularity conditions, it follows that $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$; for serially uncorrelated errors, see White (1980) and White (1984), and for serially correlated errors, see Newey and West (1987) and Andrews (1991b). By the continuous mapping theorem and the fact that the selection matrix is non-random, it follows that $\hat{\mathbf{Q}}_m \xrightarrow{p} \mathbf{Q}_m$, $\hat{\mathbf{\Omega}}_{m,p} \xrightarrow{p} \mathbf{\Omega}_{m,p}$, and $\hat{\mathbf{a}}_m \xrightarrow{p} \mathbf{a}_m$.

We now consider the estimator for the local parameter $\boldsymbol{\delta}$. Unlike $\mathbf{D}_{\boldsymbol{\theta}_m}$, $\mathbf{Q}_m$, $\mathbf{\Omega}_{m,p}$, and $\mathbf{a}_m$, there is no consistent estimator for the parameter $\boldsymbol{\delta}$. This implies that the optimal weights cannot be estimated consistently. We propose to use the asymptotically unbiased estimator for $\boldsymbol{\delta}$. Let $\hat{\boldsymbol{\delta}} = \sqrt{n}\hat{\boldsymbol{\gamma}}$ where $\hat{\boldsymbol{\gamma}}$ are the estimates from the full model. From Lemma 1, we have

$$\hat{\boldsymbol{\delta}} = \sqrt{n}\hat{\boldsymbol{\gamma}} \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \mathbf{\Pi}_\ell\mathbf{Q}^{-1}\mathbf{R} \sim \mathbf{N}(\boldsymbol{\delta}, \mathbf{\Pi}_\ell\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{\Pi}'_\ell) \tag{4.8}$$

where $\mathbf{\Pi}_\ell = (\mathbf{0}_{\ell\times k}, \mathbf{I}_\ell)$. As shown above, $\hat{\boldsymbol{\delta}}$ is an asymptotically unbiased estimator for $\boldsymbol{\delta}$. The limiting distribution of the plug-in estimator $\hat{\boldsymbol{\delta}}$ is $\mathbf{R}_{\boldsymbol{\delta}}$ which is a linear function of the normal random vector $\mathbf{R}$. We use this result to establish the asymptotic distribution of the plug-in averaging estimator.

Note that the first term of $\zeta_{m,p}$ can be rewritten as $\mathbf{a}'_m\boldsymbol{\delta}\boldsymbol{\delta}'\mathbf{a}_p$. Hence, we can estimate $\boldsymbol{\delta}\boldsymbol{\delta}'$ instead of $\boldsymbol{\delta}$. Since $\mathbf{R}_{\boldsymbol{\delta}}\mathbf{R}'_{\boldsymbol{\delta}}$ has mean $\boldsymbol{\delta}\boldsymbol{\delta}' + \mathbf{\Pi}_\ell\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{\Pi}'_\ell$, another possible estimator is $n\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}' - \mathbf{\Pi}_\ell\hat{\mathbf{Q}}^{-1}\hat{\mathbf{\Omega}}\hat{\mathbf{Q}}^{-1}\mathbf{\Pi}'_\ell$ for $\boldsymbol{\delta}\boldsymbol{\delta}'$. However, it might happen that the estimator of the squared bias terms, the diagonal terms of $\boldsymbol{\delta}\boldsymbol{\delta}'$, are negative. Furthermore, the asymptotic distribution of the squared bias estimator is more complicated. Therefore, we only consider the estimator $\hat{\boldsymbol{\delta}}$.

The following assumption is imposed on the estimator of the covariance matrix.

Assumption 3. There exists $\hat{\mathbf{\Omega}}$ such that $\hat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$.

Assumption 3 is a high-level condition on the estimator of the covariance matrix. Rather than impose regularity conditions, we assume there exists a consistent estimator for $\mathbf{\Omega}$. The consistent estimators for the covariance matrix are given in (4.4) and (4.5) for serially uncorrelated errors and serially correlated errors, respectively. The sufficient condition for the consistency is e_i is i.i.d. or a martingale difference sequence with finite fourth moment. For serial correlation, data is a mean zero α-mixing or φ-mixing sequence.

Theorem 2. *Suppose Assumptions 1-3 hold. As $n \to \infty$, we have*

$$\mathbf{w}'\hat{\boldsymbol{\zeta}}\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$$

where $\boldsymbol{\zeta}^*$ *is an* $M \times M$ *matrix with the* (m,p)*th element*

$$\zeta^*_{m,p} = \mathbf{R}'_{\boldsymbol{\delta}}\mathbf{a}_m\mathbf{a}'_p\mathbf{R}_{\boldsymbol{\delta}} + \mathbf{D}'_{\boldsymbol{\theta}_m}\mathbf{Q}_m^{-1}\boldsymbol{\Omega}_{m,p}\mathbf{Q}_p^{-1}\mathbf{D}_{\boldsymbol{\theta}_p}$$

and $\mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \boldsymbol{\Pi}_\ell\mathbf{Q}^{-1}\mathbf{R}$*. Also, we have*

$$\hat{\mathbf{w}}_{pia} \xrightarrow{d} \mathbf{w}^*_{pia} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}, \tag{4.9}$$

and

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}_{pia}) - \mu\big) \xrightarrow{d} \sum_{m=1}^{M} w^*_{pia,m}\Lambda_m \tag{4.10}$$

where $\Lambda_m = \mathbf{a}'_m\boldsymbol{\delta} + \mathbf{b}'_m\mathbf{R}$*.*

Theorem 2 shows that the estimated weights are asymptotically random. In order to derive the asymptotic distribution of the plug-in averaging estimator, we show that there is joint convergence in distribution of all submodel estimators $\tilde{\mu}_m$ and estimated weights $\hat{\mathbf{w}}_{pia}$. The joint convergence in distribution comes from the fact that both Λ_m and $w^*_{pia,m}$ can be expressed in terms of the normal random vector $\mathbf{R}$. It turns out the limiting distribution of the plug-in averaging estimator is not normally distributed. Instead, it is a non-linear function of the normal random vector $\mathbf{R}$.

The non-normal nature of the limiting distribution of the averaging estimator with data-driven weights is also pointed out by Hjort and Claeskens (2003) and Claeskens and Hjort (2008). The result is useful to construct the confidence interval.

5 AIC, S-AIC and JMA Estimators

In this section, we present the asymptotic distributions of the AIC model selection estimator, the S-AIC model averaging estimator, and the Jackknife Model Averaging estimator. The limiting distributions of AIC, S-AIC, and JMA estimators are non-standard in the local asymptotic framework. We also present the results of the two-model case.

5.1 AIC and Smoothed AIC

The model selection estimator based on information criteria is a special case of the model averaging estimator. The model selection puts the whole weight on the model with the smallest value of the information criterion and give other models zero weight. Hence, the weight function of the model selection estimator can be described by the indicator function.

The AIC for the linear regression model (2.4) is

$$\mathrm{AIC}_m = n\log(\tilde{\sigma}^2_m) + 2(k+\ell_m), \quad m = 1, 2, ..., M,$$

where $\tilde{\sigma}_m^2 = n^{-1}\sum_{i=1}^n \tilde{e}_{mi}^2$ and $\tilde{e}_{mi}$ are the least squares residuals from the submodel m, that is, $\tilde{e}_{mi} = y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}_m - \mathbf{z}_{mi}'\tilde{\boldsymbol{\gamma}}_m$ and $\mathbf{z}_{mi} = \boldsymbol{\Pi}_m\mathbf{z}_i$. The AIC model selection estimator is thus

$$\bar{\mu}(\hat{\mathbf{w}}_{aic}) = \sum_{m=1}^M \hat{w}_{aic,m}\tilde{\mu}_m,$$
$$\hat{w}_{aic,m} = \mathbf{1}\{\mathrm{AIC}_m = \min(\mathrm{AIC}_1, \mathrm{AIC}_2, ..., \mathrm{AIC}_M)\}.$$

Instead of estimating the regression function based on a single model, the S-AIC model averaging estimator proposed by Buckland, Burnham, and Augustin (1997) assigns the weights of each candidate models by using the exponential Akaike information criterion. The weight for each submodel is proportional to the log-likelihood of model. The S-AIC model averaging estimator is defined as

$$\bar{\mu}(\hat{\mathbf{w}}_{saic}) = \sum_{m=1}^M \hat{w}_{saic,m}\tilde{\mu}_m, \tag{5.1}$$
$$\hat{w}_{saic,m} = \frac{\exp(-\frac{1}{2}\mathrm{AIC}_m)}{\sum_{m=1}^M \exp(-\frac{1}{2}\mathrm{AIC}_m)}. \tag{5.2}$$

The S-AIC weight is similar to the smoothed Bayesian information criterion (S-BIC) model averaging where the weights are chosen by using the exponential Bayesian information criterion. The S-BIC weight is $\exp(-\frac{1}{2}\mathrm{BIC}_m)/\sum_{m=1}^M \exp(-\frac{1}{2}\mathrm{BIC}_m)$, where $\mathrm{BIC}_m = n\log(\tilde{\sigma}_m^2) + \log(n)(k+\ell_m)$. The weights of the Bayesian model averaging are interpreted as the posterior model probabilities. Therefore, the S-AIC weight may be interpreted as the model probability.

The S-AIC model averaging estimator is appealing because of its simplicity. Also, there is a closed form expression of the S-AIC weights for any number of submodels. However, both AIC and S-AIC are not robust for heteroskedastic regressions. The misspecification-robust version of AIC is Takeuchi information criterion, see Burnham and Anderson (2002). Furthermore, the S-AIC weights ignore the covariances between the submodel estimators. Also, the S-AIC weights are formed based on the global fit of the model, and the weights does not adjust according to the parameter of interest.

Hjort and Claeskens (2003) and Claeskens and Hjort (2008) show the limiting distributions of the AIC model selection estimator and the S-AIC model averaging estimator in the likelihood framework. Let $\mathrm{AIC}_\emptyset$ be the AIC for the narrow model. Following Theorem 5.4 of Claeskens and Hjort (2008), we can show that the

$$\mathrm{AIC}_\emptyset - \mathrm{AIC}_m \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_m\mathbf{R}_{\boldsymbol{\delta}} - 2(k+\ell_m) \tag{5.3}$$

where $\boldsymbol{\Sigma}_m = \mathbf{V}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Pi}_m'\left(\boldsymbol{\Pi}_m\mathbf{V}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Pi}_m'\right)^{-1}\boldsymbol{\Pi}_m\mathbf{V}_{\boldsymbol{\delta}}^{-1}$ and $\mathbf{V}_{\boldsymbol{\delta}} = \boldsymbol{\Pi}_\ell\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}\boldsymbol{\Pi}_\ell'$.

Note that (5.3) can be expressed as $\mathbf{G}'\boldsymbol{\Psi}_m\mathbf{G} - 2(k+\ell_m)$ where $\mathbf{G} \sim \mathbf{N}(\mathbf{V}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{\delta}, \mathbf{I}_\ell)$ and $\boldsymbol{\Psi}_m = \mathbf{V}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{\Pi}_m'\left(\boldsymbol{\Pi}_m\mathbf{V}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Pi}_m'\right)^{-1}\boldsymbol{\Pi}_m\mathbf{V}_{\boldsymbol{\delta}}^{-1/2}$. Here $\mathbf{G}'\boldsymbol{\Psi}_m\mathbf{G}$ has a noncentral chi-squared distribution with ℓ_m degrees of freedom and non-centrality parameter $\lambda_m = \boldsymbol{\delta}'\mathbf{V}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{\Psi}_m\mathbf{V}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{\delta}$. Similar to the plug-in averaging estimator, the asymptotic distributions of the AIC model selection estimator and the

S-AIC model averaging estimator can be expressed as a non-linear functions of the normal random vector $\mathbf{R}$.

Theorem 3. *Suppose Assumptions 1-2 hold. As $n \to \infty$, the asymptotic distribution of the S-AIC model averaging estimator is*

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}_{saic}) - \mu\big) \xrightarrow{d} \sum_{m=1}^{M} w^*_{saic,m}\Lambda_m$$

where

$$w^*_{saic,m} = \frac{\exp(\frac{1}{2}\mathbf{R}'_{\boldsymbol{\delta}}\boldsymbol{\Sigma}_m\mathbf{R}_{\boldsymbol{\delta}} - (k+\ell_m))}{\sum_{m=1}^{M}\exp(\frac{1}{2}\mathbf{R}'_{\boldsymbol{\delta}}\boldsymbol{\Sigma}_m\mathbf{R}_{\boldsymbol{\delta}} - (k+\ell_m))}$$

and $\Lambda_m = \mathbf{a}'_m\boldsymbol{\delta} + \mathbf{b}'_m\mathbf{R}$.

5.2 Jackknife Model Averaging Estimator

The Jackknife Model Averaging estimator is proposed by Hansen and Racine (2012). They suggest to select the weights by minimizing a leave-one-out cross-validation criterion. They show the asymptotic optimality of the JMA estimator. That is, the average squared error of the JMA estimator is asymptotic equivalent to the lowest expected squared error. The asymptotic optimality of the cross-validation criterion is first established by Li (1987) for model selection in homoskedastic regression with an infinite number of regressors. Following Li (1987), Andrews (1991a) shows the asymptotic optimality of the cross-validation criterion for model selection for heteroskedastic regressions. Hansen and Racine (2012) extend the asymptotic optimality from model selection to model averaging. However, the optimality result of Theorem 1 in Hansen and Racine (2012) requires the condition which there is no submodel m for which the bias is zero. Therefore, it cannot apply to the context of the linear regression model with a finite number of regressors. In other words, the JMA is not asymptotically optimal in our framework.

Define the leave-one-out cross-validation criterion for the averaging estimator for the linear regression model (2.4) as follows:

$$\mathrm{CV}_n(\mathbf{w}) = \frac{1}{n}\mathbf{w}'\tilde{\mathbf{e}}'_{-i}\tilde{\mathbf{e}}_{-i}\mathbf{w} \tag{5.4}$$

where $\tilde{\mathbf{e}}_{-i} = (\tilde{\mathbf{e}}_{1,-i}, ..., \tilde{\mathbf{e}}_{M,-i})$ is a $n \times M$ matrix of leave-one-out least-squares residuals and $\tilde{\mathbf{e}}_{m,-i}$ are the residuals of submodel m obtained by least-squares estimation without the $i'th$ observation. The weight vector of the JMA estimator is the value which minimizes $\mathrm{CV}_n(\mathbf{w})$.

By adding and subtracting the sum of squared residuals of the full model $\frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}}$, we can rewrite (5.4) as

$$\mathrm{CV}_n(\mathbf{w}) = \frac{1}{n}\mathbf{w}'\boldsymbol{\xi}_n\mathbf{w} + \frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}} \tag{5.5}$$

where $\boldsymbol{\xi}_n$ is an $M \times M$ matrix with the (m,p)th element

$$\xi_{m,p} = \tilde{\mathbf{e}}'_{m,-i}\tilde{\mathbf{e}}_{p,-i} - \hat{\mathbf{e}}'\hat{\mathbf{e}}. \tag{5.6}$$

Note that minimizing $\mathrm{CV}_n(\mathbf{w})$ over $\mathbf{w} = (w_1, ..., w_M)$ is equivalent to minimizing $\mathbf{w}'\boldsymbol{\xi}_n\mathbf{w}$ since $\frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is not related to the weight vector $\mathbf{w}$. In the following theorem, we show that $\xi_{m,p}$ converges to a non-linear function of the normal random vector $\mathbf{R}$. The JMA estimator can be represented as

$$\bar{\mu}(\hat{\mathbf{w}}_{jma}) = \sum_{m=1}^{M} \hat{w}_{jma,m}\tilde{\mu}_m, \tag{5.7}$$

$$\hat{\mathbf{w}}_{jma} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\xi}_n\mathbf{w}. \tag{5.8}$$

Here, the weight vector is defined as the minimizer of the quadratic function of $\mathbf{w}$ which can be found by quadratic programming as the optimal fixed-weight vector and the plug-in weight vector. However, unlike the plug-in averaging estimator where the weights are tailored to the parameter of interest, the JMA estimator selects the weights based on the conditional mean function. One disadvantage of the JMA estimator is the computational burden, which is substantial when both the sample size and the number of regressors are large.

The following assumption is imposed on the data generating process.

Assumption 4. (a) $\{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1, ..., n\}$ are i.i.d. (b) $\mathrm{E}(e_i^4) < \infty$, $\mathrm{E}(x_{ji}^4) < \infty$ for $j = 1, ..., k$, and $\mathrm{E}(z_{ji}^4) < \infty$ for $j = 1, ..., \ell$.

Condition (a) in Assumption 4 is the i.i.d. assumption, which is also made in Hansen and Racine (2012). The result in Theorem 4 can be extended to the stationary case. Condition (b) is the standard assumption for the linear regression model. Note that Assumption 4 implies Assumption 2. Therefore, the results in Lemma 1, Lemma 2, and Theorem 1 hold under Assumptions 1 and 4.

Theorem 4. *Suppose Assumptions 1 and 4 hold. As $n \to \infty$, we have*

$$\mathbf{w}'\boldsymbol{\xi}_n\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}$$

where $\boldsymbol{\xi}^*$ is an $M \times M$ matrix with the (m,p)th element

$$\xi^*_{m,p} = \ddot{\mathbf{R}}'_m\mathbf{Q}\ddot{\mathbf{R}}_p + tr\left(\mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\right) + tr\left(\mathbf{Q}_p^{-1}\boldsymbol{\Omega}_p\right) \tag{5.9}$$

and $\ddot{\mathbf{R}}_m = \ddot{\mathbf{A}}_m\boldsymbol{\delta} + \ddot{\mathbf{B}}_m\mathbf{R}$ with

$$\ddot{\mathbf{A}}_m = \left(\boldsymbol{\Pi}'_\ell - \mathbf{S}_m\mathbf{Q}_m^{-1}\begin{pmatrix}\mathbf{Q}_{\mathbf{xz}} \\ \boldsymbol{\Pi}_m\mathbf{Q}_{\mathbf{zz}}\end{pmatrix}\right)\left(\mathbf{I}_\ell - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right),$$

and

$$\ddot{\mathbf{B}}_m = \left(\mathbf{Q}^{-1} - \mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}'_m\right).$$

Also, we have

$$\hat{\mathbf{w}}_{jma} \xrightarrow{d} \mathbf{w}^*_{jma} = \underset{\mathbf{w}\in\mathcal{H}_n}{\operatorname{argmin}}\, \mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}, \tag{5.10}$$

and

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}_{jma}) - \mu\big) \xrightarrow{d} \sum_{m=1}^{M} w^*_{jma,m}\Lambda_m \tag{5.11}$$

where $\Lambda_m = \mathbf{a}'_m\boldsymbol{\delta} + \mathbf{b}'_m\mathbf{R}$.

5.3 Model Averaging for the Two-Model Case

In this section, we concentrate on a special case with only two candidate models. As we mentioned in previous section, we have a closed-form solution for the weight vector when the number of total models equals two. Pötscher (2006) also analyzes the asymptotic distribution of the averaging estimator for the two-model case, but assumes the error term is normal distributed. Here, we generalize his results by relaxing the assumption on the error term and also considering the case of two non-nested candidate models.

Suppose the auxiliary regressors are partition as $\mathbf{Z} = (\mathbf{Z}\boldsymbol{\Pi}'_1, \mathbf{Z}\boldsymbol{\Pi}'_2) = (\mathbf{Z}_1, \mathbf{Z}_2)$ where $\boldsymbol{\Pi}_1 = (\mathbf{I}_{\ell_1}, \mathbf{0}_{\ell_1\times\ell_2})$ and $\boldsymbol{\Pi}_2 = (\mathbf{0}_{\ell_2\times\ell_1}, \mathbf{I}_{\ell_2})$. Then the regression model (2.4) can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \mathbf{e} \tag{5.12}$$

where $\boldsymbol{\gamma}_1$ is $\ell_1 \times 1$, $\boldsymbol{\gamma}_2$ is $\ell_2 \times 1$, and $\ell_1 + \ell_2 = \ell$. We assume the Model 1 includes the regressors $\mathbf{X}$ and $\mathbf{Z}_1$ while the Model 2 includes the regressors $\mathbf{X}$ and $\mathbf{Z}_2$. If $\ell_2 = \ell$, then the Model 1 is the restricted model and the Model 2 is the unrestricted model, which is the framework of Pötscher (2006). If $\ell_1 > 0$ and $\ell_2 > 0$, then the Model 1 and 2 are two non-nested models.

We denote the estimators of the fucus parameter for the two candidate models by $\tilde{\mu}_1 = \mu(\tilde{\boldsymbol{\theta}}_1) = \mu(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\gamma}}_1, \mathbf{0})$ and $\tilde{\mu}_2 = \mu(\tilde{\boldsymbol{\theta}}_2) = \mu(\tilde{\boldsymbol{\beta}}_2, \mathbf{0}, \tilde{\boldsymbol{\gamma}}_2)$, respectively. Let w be the weight for $\tilde{\mu}_1$ and $1 - w$ be the weight for $\tilde{\mu}_2$. The averaging estimator for the two-model case is

$$\bar{\mu}(w) = w\tilde{\mu}_1 + (1 - w)\tilde{\mu}_2. \tag{5.13}$$

Let w^o be the infeasible optimal fixed-weight. The following corollary describes the AMSE of the averaging estimator with the infeasible optimal fixed-weight.

Corollary 1. *Suppose Assumptions 1-2 hold. Then the AMSE of the averaging estimator for the two-model case is*

$$\text{AMSE}(\bar{\mu}(w)) = w^2\zeta_{1,1} + (1 - w)^2\zeta_{2,2} + 2w(1 - w)\zeta_{1,2}$$

where $\zeta_{m,p}$ is defined in (3.2). The weight w which minimizes $\mathrm{AMSE}(\bar{\mu}(w))$ *is*

$$w^o = \begin{cases} \frac{\zeta_{2,2}-\zeta_{1,2}}{\zeta_{1,1}+\zeta_{2,2}-2\zeta_{1,2}} & \text{if } \zeta_{1,2} < \min\{\zeta_{1,1}, \zeta_{2,2}\}, \\ 1 & \text{if } \zeta_{1,1} \le \zeta_{1,2} < \zeta_{2,2}, \\ 0 & \text{if } \zeta_{2,2} \le \zeta_{1,2} < \zeta_{1,1}, \end{cases}$$

and the minimized AMSE is

$$\mathrm{AMSE}(\bar{\mu}(w^o)) = \begin{cases} \frac{\zeta_{1,1}\zeta_{2,2}-\zeta_{1,2}^2}{\zeta_{1,1}+\zeta_{2,2}-2\zeta_{1,2}} & \text{if } \zeta_{1,2} < \min\{\zeta_{1,1}, \zeta_{2,2}\}, \\ \zeta_{1,1} & \text{if } \zeta_{1,1} \le \zeta_{1,2} < \zeta_{2,2}, \\ \zeta_{2,2} & \text{if } \zeta_{2,2} \le \zeta_{1,2} < \zeta_{1,1}. \end{cases}$$

The values of $\zeta_{1,1}$ and $\zeta_{2,2}$ in Corollary 1 represent the AMSE of the Model 1 and 2, respectively. As long as $\zeta_{1,2} < min\{\zeta_{1,1}, \zeta_{2,2}\}$, the AMSE of the averaging estimator with the optimal fixed-weight is strictly less than the AMSE of any convex combination of the Model 1 and 2.

We now consider the averaging estimator with data-driven weights when there are only two candidate models. Let $\hat{w}_{saic}$, $\hat{w}_{pia}$ and $\hat{w}_{jma}$ be the weights chosen by the S-AIC model averaging estimator, the plug-in averaging estimator, and the JMA estimator. From Theorem 3, it can be shown that the AMSE of the S-AIC model averaging estimator $\bar{\mu}(\hat{w}_{siac})$ is

$$AMSE\big(\bar{\mu}(\hat{w}_{saic})\big) = \mathrm{E}\Big(w_{saic}^{*2}\zeta_{1,1} + \big(1 - w_{saic}^*\big)^2\zeta_{2,2} + 2w_{saic}^*\big(1 - w_{saic}^*\big)\zeta_{1,2}\Big)$$

where $w_{saic}^* = (\exp(2^{-1}\mathbf{R}_{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1\mathbf{R}_{\boldsymbol{\delta}} - (k+\ell_1)))/(\sum_{m=1}^2 \exp(2^{-1}\mathbf{R}_{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_m\mathbf{R}_{\boldsymbol{\delta}} - (k+\ell_m)))$. The following corollary presents the AMSE of the plug-in averaging estimator and the JMA estimator.

Corollary 2. *(a) Suppose Assumptions 1-3 hold. Then the AMSE of the plug-in averaging estimator for the two-model case is* $\mathrm{AMSE}\big(\bar{\mu}(\hat{w}_{pia})\big) = E\big(w_{pia}^{*2}\zeta_{1,1} + \big(1 - w_{pia}^*\big)^2\zeta_{2,2} + 2w_{pia}^*\big(1 - w_{pia}^*\big)\zeta_{1,2}\big)$ *where*

$$w_{pia}^* = \begin{cases} \frac{\zeta_{2,2}^*-\zeta_{1,2}^*}{\zeta_{1,1}^*+\zeta_{2,2}^*-2\zeta_{1,2}^*} & \text{if } \zeta_{1,2}^* < \min\{\zeta_{1,1}^*, \zeta_{2,2}^*\}, \\ 1 & \text{if } \zeta_{1,1}^* \le \zeta_{1,2}^* < \zeta_{2,2}^*, \\ 0 & \text{if } \zeta_{2,2}^* \le \zeta_{1,2}^* < \zeta_{1,1}^*, \end{cases}$$

and $\zeta_{m,p}^*$ is defined in Theorem 2.

(b) Suppose Assumptions 1 and 4 hold. Then the AMSE of the Jackknife Model Averaging estimator for the two-model case is $\mathrm{AMSE}\big(\bar{\mu}(\hat{w}_{jma})\big) = E\big(w_{jma}^{*2}\xi_{1,1} + \big(1 - w_{jma}^*\big)^2\xi_{2,2} + 2w_{jma}^*\big(1 - w_{jma}^*\big)\xi_{1,2}\big)$ *where*

$$w_{jma}^* = \begin{cases} \frac{\xi_{2,2}^*-\xi_{1,2}^*}{\xi_{1,1}^*+\xi_{2,2}^*-2\xi_{1,2}^*} & \text{if } \xi_{1,2}^* < \min\{\xi_{1,1}^*, \xi_{2,2}^*\}, \\ 1 & \text{if } \xi_{1,1}^* \le \xi_{1,2}^* < \xi_{2,2}^*, \\ 0 & \text{if } \xi_{2,2}^* \le \xi_{1,2}^* < \xi_{1,1}^*, \end{cases}$$

and $\xi_{m,p}^*$ is defined in Theorem 4.

Note that w^o, w^*_{pia}, and w^*_{jma} have the similar form but different interpretations. w^o is non-random since all $\zeta_{1,1}$, $\zeta_{2,2}$, and $\zeta_{1,2}$ are constants. Both w^*_{pia} and w^*_{jma} are random because $\zeta^*_{m,p}$ and $\xi^*_{m,p}$ are a non-linear function of the normal random vector $\mathbf{R}$. The results also implies the non-standard limiting distribution of the data-driven estimator in the simple two-model case.

6 Simulation Results

In this section, we investigate the finite sample mean square error of the plug-in averaging estimator via Monte Carlo experiments.

6.1 Simulation Setup

We consider a linear regression model with a finite number of regressors

$$y_i = \sum_{j=1}^{J} \theta_j x_{ji} + e_i, \quad i = 1, ..., n. \tag{6.1}$$

We let x_{1i} and x_{2i} be the core regressors and the remaining x_{ji} are the auxiliary regressors. We set $x_{1i} = 1$ to be the intercept. The random variables $(x_{2i}, ..., x_{Ji})'$ are generated from a joint normal distribution $N(0, \mathbf{\Sigma})$ where the diagonal elements of $\mathbf{\Sigma}$ are 1, $\mathrm{E}(x_{2i}x_{ji}) = \rho_1$ for $j \geq 3$, and $\mathrm{E}(x_{ji}x_{ki}) = \rho_2$ for $j, k \geq 3$ and $j \neq k$. The error term e_i is generated from a normal distribution $N(0, \sigma_i^2)$, where $\sigma_i^2 = 1$ for the homoskedastic simulation and $\sigma_i^2 = x_{2i}^2$ for the heteroskedastic simulation.

The parameters are determined by the following two rules:

$$\mathrm{DGP}_1 : \boldsymbol{\theta} = \left(\frac{\sqrt{n}}{8}, \frac{\sqrt{n}}{8}, 1, \frac{\ell - 1}{\ell}, ..., \frac{1}{\ell} \right)' c/\sqrt{n}, \tag{6.2}$$

$$\mathrm{DGP}_2 : \boldsymbol{\theta} = \left(-\frac{\sqrt{n}}{8}, \frac{\sqrt{n}}{8}, -1, \frac{\ell - 1}{\ell}, ..., -\frac{1}{\ell} \right)' c/\sqrt{n}, \tag{6.3}$$

where $\ell = J - 2$. The parameter c is selected to control the population $R^2 = \boldsymbol{\theta}_2' \mathbf{\Sigma} \boldsymbol{\theta}_2 / (1 + \boldsymbol{\theta}_2' \mathbf{\Sigma} \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2 = (\theta_2, ..., \theta_J)'$ and R^2 varies on a grid between 0.1 and 0.9. The local parameters are determined by $\delta_j = \sqrt{n}\theta_j = c(\ell - j + 3)/\ell$ for $j \geq 3$. The number of the regressors is varied between $J = 3$, 5, 7, and 9. We consider all possible submodels, that is, the number of models is $M = 2^{J-2}$.

6.2 Finite Sample Comparison

We consider six estimators: (1) AIC model selection estimator (labeled AIC), (2) BIC model selection estimator (labeled BIC), (3) S-AIC model averaging estimator (labeled S-AIC), (4) S-BIC model averaging estimator (labeled S-BIC), (5) Jackknife Model Averaging estimator (labeled JMA), and (6) Plug-In averaging estimator (labeled Plug-In). The parameter of interest is $\mu = \theta_2$. To evaluate the finite behavior of the averaging estimators, we compute the risk based on the quadratic loss function, i.e. $\mathrm{E}\big(n(\hat{\theta}_2 - \theta_2)^2\big)$. The risk (expected squared error) is calculated

by averaging across $5,000$ random samples. We normalize the risk by dividing by the optimal asymptotic risk. The optimal asymptotic risk is defined as $\mathbf{w}^{o\prime}\boldsymbol{\zeta}\mathbf{w}^{o}$, where $\boldsymbol{\zeta}$ and $\mathbf{w}^{o}$ are defined in (3.2) and (3.3). The sample sizes are 50, 100, 150, 200 for $M = 2$, 8, 32, and 128.

Figures 1 and 2 show the risk functions for DGP_1 and DGP_2 with $(\rho_1, \rho_2) = (0.3, 0.1)$ in the homoskedastic simulation and Figures 3 and 4 show the risk functions for DGP_1 with $(\rho_1, \rho_2) = (0.3, 0.1)$ and $(0.6, 0.4)$ in the heteroskedastic simulation.[1] In each figure, the risk is displayed for $M = 2$, 8, 32, and 128, respectively. The dotted line represents the AIC model selection estimator, the solid line with asterisk represents the BIC model selection estimator, the dash-dotted line represents the S-AIC model averaging estimator, the dash line with circle represents the S-BIC model averaging estimator, the dashed lines represents the JMA estimator, and the solid line represents the plug-in averaging estimator.

There are several remarks about the simulations results. First, the risk of all estimators increases as the number of models increases. When we only consider the restricted and nonrestricted models, i.e. $M = 2$, all estimators have similar risk. Second, it can be seen that the plug-in averaging estimator dominates other estimators in most ranges of the population R^2. The JMA estimator has smaller risk than the S-AIC estimator for DGP_2, but S-AIC achieves lower risk when M and R^2 are larger for DGP_1. The S-BIC estimator and the BIC model selection estimator have poor performance relative to the other methods in most cases. Also note that the model-averaging-type estimators have lower risk than the model-selection-type counterpart estimators. Third, all estimators have smaller normalized risk under heteroskedastic errors, but the ranking of the estimators in the heteroskedastic simulation is quite similar to that in the homoskedastic simulation. Fourth, the normalized risk of the plug-in estimator is close to 1 for DGP_1, meaning that it is close to that of the averaging estimator with infeasible optimal fixed weights. The normalized risk of the plug-in estimator is getting far from 1 as the number of models increases for DGP_2. Also note that the risk of all estimators has smaller variation across the parameters R^2 in DGP_2 than those in DGP_1. Fifth, as ρ_1 and ρ_2 increase, the risk of all estimators increases. However, the ranking of the estimators for $(\rho_1, \rho_2) = (0.6, 0.4)$ is quite similar to that for $(\rho_1, \rho_2) = (0.3, 0.1)$.

Tables 1 and 2 report the maximum risk and maximum regret of the estimators. Here we define the regret as the difference between the risk of the estimator and the optimal asymptotic risk (labeled Opt). The maximum regret is the largest value of the regret across the parameters R^2. The maximum risk is defined as the same way. It is clear that the plug-in averaging estimator achieves the minimax risk and minimax regret in all simulation cases. One interesting observation from Tables 1 and 2 is that the results between DGP_1 and DGP_2 are quite different. The maximum risk of the averaging estimator with infeasible optimal fixed weights increases as the number of models increases for DGP_1, but decreases as the number of models increases for DGP_2. Unlike other estimators, the plug-in averaging estimator has relatively low maximum regret for DGP_1. Also note that the maximum risk/regert of all data-driven estimators are close to each other for DGP_2. Another interesting observation is that all estimators have larger maximum risk but smaller

[1] We report the results of the heteroskedastic simulations for DGP_1 only for space considerations. All results are available on request from the author.

maximum regret in the heteroskedastic simulation than in the homoskedastic simulation.

6.3 Robust Simulation

We consider two extended setups to investigate the finite sample behavior of the plug-in averaging estimator. The data generating process is based on (6.1) with $(\rho_1, \rho_2) = (0.3, 0.1)$ and the parameters are determined by the following:

$$\text{DGP}_3 : \boldsymbol{\theta} = \left(-\frac{\sqrt{n}}{8}, \frac{\sqrt{n}}{8}, \left(-1, \frac{\ell-1}{\ell}, ..., -\frac{1}{\ell}\right)^a\right)' c/\sqrt{n}, \tag{6.4}$$

$$\text{DGP}_4 : \boldsymbol{\theta} = \left(-\frac{\sqrt{n}}{b}, \frac{\sqrt{n}}{b}, -1, \frac{\ell-1}{\ell}, ..., -\frac{1}{\ell}\right)' c/\sqrt{n}, \tag{6.5}$$

where $\ell = 5$, $a = \{0.5, 1, 1.5, 2\}$, $b = \{4, 6, 8, 10\}$, and c is selected to control the population R^2. The sample size is 150. The number of simulations is 5000.

Figures 5 and 6 show the risk functions for DGP_3 and DGP_4, respectively. From Figure 5, it can be seen that the magnitude of risk decreases as the parameter a increases. This implies that when the coefficients of auxiliary regressors decline more quickly, i.e. a is larger, the risk of all estimators are getting closer. Figure 6 shows the S-AIC, S-BIC, and JMA estimators achieves lower risk than the plug-in averaging estimator when the parameter b and R^2 are small. This implies that when the auxiliary regressors have a greater influence on the model, i.e. b is larger, the plug-in averaging estimator performs better than other averaging estimators. Table 3 reports the maximum risk and maximum regret for DGP_3 and DGP_4. It shows that the plug-in averaging estimator still achieves the minimax risk and minimax regret across the parameters a, b, and R^2, even if the plug-in averaging estimator has larger risk in some ranges of the population R^2 displayed in Figures 5 and 6.

7 Confidence Intervals

In this section, we propose a plug-in method to construct the confidence interval for the focus parameter μ. Since μ is a scalar, the t-statistic is used to construct the confidence interval. Define

$$\hat{V} = \sum_{m=1}^{M} \hat{w}_m^2 \hat{\mathbf{D}}'_{\boldsymbol{\theta}_m} \hat{\mathbf{Q}}_m^{-1} \hat{\boldsymbol{\Omega}}_m \hat{\mathbf{Q}}_m^{-1} \hat{\mathbf{D}}_{\boldsymbol{\theta}_m} + 2 \sum_{m<p} \hat{w}_m \hat{w}_p \hat{\mathbf{D}}'_{\boldsymbol{\theta}_m} \hat{\mathbf{Q}}_m^{-1} \hat{\boldsymbol{\Omega}}_{m,p} \hat{\mathbf{Q}}_p^{-1} \hat{\mathbf{D}}_{\boldsymbol{\theta}_p}, \tag{7.1}$$

where $\hat{w}_m$ could be the weight chosen by the plug-in averaging estimator, or other averaging estimators with data-driven weights. The model averaging t-statistic for μ is

$$t_n(\mu) = \frac{\bar{\mu}(\hat{\mathbf{w}}) - \mu}{\widehat{se}_n} \tag{7.2}$$

where $\bar{\mu}(\hat{\mathbf{w}})$ is the averaging estimators with data-driven weights $\hat{\mathbf{w}}$ and $\widehat{se}_n = (\hat{V}/n)^{1/2}$.

Theorem 5. *Suppose Assumptions 1, 3, and 4 hold. As $n \to \infty$, we have*

$$t_n(\mu) \xrightarrow{d} (V^*)^{-1/2} \sum_{m=1}^{M} w_m^* \Lambda_m$$

where $V^* = \sum_{m=1}^{M} w_m^{*2} \mathbf{D}_{\boldsymbol{\theta}_m}' \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_m \mathbf{Q}_m^{-1} \mathbf{D}_{\boldsymbol{\theta}_m} + 2 \sum_{m<p} w_m^* w_p^* \mathbf{D}_{\boldsymbol{\theta}_m}' \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_{m,p} \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p}$ *and* $\Lambda_m = \mathbf{a}_m' \boldsymbol{\delta} + \mathbf{b}_m' \mathbf{R}$.

Theorem 5 is a general statement for all averaging estimators with data-driven weights. For example, if weights are chosen by the plug-in averaging estimator, then $w_m^* = w_{pia,m}^*$, where $w_{pia,m}^*$ is defined in Theorem 2. Theorem 5 states that the asymptotic distribution of the model averaging t-statistic is not normally distributed. Instead, it is characterized by a non-linear function of the normal random vector $\mathbf{R}$.

Let $\mathrm{CI}_n(\alpha)$ denote the $1-\alpha$ percent confidence interval for parameter μ where α is the nominal size. By inverting the t-statistic, we construct the confidence interval with the nominal level $1-\alpha$ for the focus parameter μ as $\mathrm{CI}_n(\alpha) = \{\mu : t_n(\mu) \leq c_{n,1-\alpha}\}$ where $c_{n,1-\alpha}$ is the critical value. The naive way to construct the confidence interval is to use the $1-\alpha$ quantile of the standard normal distribution as the critical value. For a standard two-sided symmetric confidence interval, the naive confidence interval is defined as

$$\mathrm{CI}_{1n}(\alpha) = [\bar{\mu}(\hat{\mathbf{w}}) - z_{1-\alpha/2} \widehat{se}_n,\ \bar{\mu}(\hat{\mathbf{w}}) + z_{1-\alpha/2} \widehat{se}_n] \tag{7.3}$$

where $z_{1-\alpha/2}$ is $1-\alpha/2$ quantile of the standard normal distribution. The naive confidence interval based on normal approximations is easily to implement, but it is not a valid method since $t_n(\mu)$ is not normally distributed.

Buckland, Burnham, and Augustin (1997) propose a modified confidence interval which addresses the uncertainty involved in the model selection/averaging step. They assume perfect correlation between any two models, which leads to a simplified formula for the variance. The confidence interval suggested by Buckland, Burnham, and Augustin (1997) is defined as

$$\mathrm{CI}_{2n}(\alpha) = [\bar{\mu}(\hat{\mathbf{w}}) - z_{1-\alpha/2} \widetilde{se}_n,\ \bar{\mu}(\hat{\mathbf{w}}) + z_{1-\alpha/2} \widetilde{se}_n] \tag{7.4}$$

where $\widetilde{se}_n = \sum_{m=1}^{M} w_m (\tilde{\sigma}_m^2/n + (\tilde{\mu}_m - \bar{\mu}(\hat{\mathbf{w}}))^2)^{1/2}$ and $\tilde{\sigma}_m^2 = \hat{\mathbf{D}}_{\boldsymbol{\theta}_m}' \hat{\mathbf{Q}}_m^{-1} \hat{\boldsymbol{\Omega}}_m \hat{\mathbf{Q}}_m^{-1} \hat{\mathbf{D}}_{\boldsymbol{\theta}_m}$. Here we do not need to estimate the covariance between any two submodels to calculate the standard error $\widetilde{se}_n$. However, the confidence interval proposed by Buckland, Burnham, and Augustin (1997) may still have incorrect coverage probabilities due to the non-standard distribution of the model averaging t-statistic.

A straightforward way to construct the confidence interval with the correct coverage probabilities is to set the critical value as the $1-\alpha$ quantile of the asymptotic distribution derived in Theorem 5. However, this quantile depends on unknown local parameters $\boldsymbol{\delta}$, and $\boldsymbol{\delta}$ cannot be estimated consistently. This implies the quantile cannot be estimated consistently as well. Here we propose a plug-in method to construct the confidence interval. We first estimate the full model

and obtain the estimators $\hat{\boldsymbol{\delta}}$, $\hat{\mathbf{Q}}$, $\hat{\boldsymbol{\Omega}}$, and $\hat{\mathbf{D}}_{\boldsymbol{\theta}}$. Second, we calculate the data-driven weights and estimate the standard error based on (7.1). Third, we simulate the asymptotic distribution derived in Theorem 5 based on the plug-in estimators $\hat{\boldsymbol{\delta}}$, $\hat{\mathbf{Q}}$, $\hat{\boldsymbol{\Omega}}$, and $\hat{\mathbf{D}}_{\boldsymbol{\theta}}$. Then we set the critical value as the $1-\alpha$ quantile from the simulation. Therefore, the plug-in symmetric two-sided confidence interval is defined as

$$\mathrm{CI}_{3n}(\alpha) = [\bar{\mu}(\hat{\mathbf{w}}) - \hat{c}_{n,1-\alpha}\widehat{se}_n,\ \bar{\mu}(\hat{\mathbf{w}}) + \hat{c}_{n,1-\alpha}\widehat{se}_n] \tag{7.5}$$

where $\hat{c}_{n,1-\alpha}$ is the $1-\alpha$ quantile of the simulated distribution.

7.1 Asymptotic Quantiles

As pointed out in Theorem 5, the asymptotic distribution of the model averaging t-statistic is non-standard. Figures 7 and 8 show the quantile functions of the model averaging t-statistics for DGP_1 and DGP_2 under homoskedastic errors. We set $\alpha = 0.05$. We simulate the asymptotic distribution and compute the quantile function based on Theorem 5. The quantile function is approximated by using $5,000$ random samples. The parameter of interest is $\mu = \theta_2$ and the weights are chosen by the plug-in averaging estimator.

In each figure, the quantile functions are displayed for $M = 2$, 8, 32, and 128, respectively. The dashed lines represents the quantile function for $(\rho_1, \rho_2) = (0.75, 0.75)$, the dash-dotted lines represents the quantile function for $(\rho_1, \rho_2) = (0.5, 0.5)$, the dotted lines represents the quantile function for $(\rho_1, \rho_2) = (0.25, 0.25)$, and the solid line represents the quantile function based on the standard normal distribution.

The behavior of the quantile functions are quite similar across different number of the models. It can be seen that the asymptotic quantiles of the t-statistics are far from those of the standard normal distribution. This implies the confidence intervals using $(-1.96, 1.96)$, the 95% quantile of the standard normal distribution, as critical points have incorrect coverage probabilities. Also note that the asymptotic quantile increases as ρ_1 and ρ_2 increase. One interesting observation from Figures 7 and 8 is that the quantile functions are asymmetric. For DGP_1, we have larger upper critical values, while for DGP_2, we have smaller lower critical values.

7.2 Coverage Probabilities

We now compare the coverage probabilities of the following methods: (1) Naive confidence interval (labeled Naive), (2) Buckland, Burnham, and Augustin (1997)' confidence interval (labeled BBA), (3) Plug-In confidence interval (labeled Plug-In). The finite-sample coverage probabilities of the nominal 90% and 95% symmetric two-sided confidence intervals for DGP_1 and DGP_2 under homoskedastic errors with $(\rho_1, \rho_2) = (0.75, 0.75)$ are reported in Table 4. The parameter of interest is $\mu = \theta_2$ and the weights are chosen by the plug-in averaging estimator. The number of repetition is 1,000. For the plug-in confidence interval, the critical value is approximated by simulation using $1,000$ random samples.

As we expected, the coverage probabilities of the Naive confidence intervals are lower than the nominal level 90% and 95%. The Buckland, Burnham, and Augustin (1997)' confidence intervals have better performance than the naive confidence intervals, however in some cases, the coverage probabilities of the Buckland, Burnham, and Augustin (1997)' confidence intervals are larger than the nominal values. The plug-in confidence intervals have the best performance among the three methods, and the coverage probabilities of the plug-in confidence intervals are quite close to the nominal values.

8 An Empirical Example

In this section, we apply the plug-in model averaging method to cross-country growth regressions. The challenge of empirical research on economic growth is that one does not know exactly what explanatory variables should be included in the true model. Many studies attempt to identify the variables explaining the differences in growth rates across countries by regressing the average growth rate of GDP per capita on a large set of potentially relevant variables, see Durlauf, Johnson, and Temple (2005) for a literature review. Due to the limited number of the observations and a large amount of the candidate variables, the empirical growth literature has been heavily criticized for its kitchen-sink approach.

In order to take into account the model uncertainty, Bayesian model averaging techniques have been applied to empirical growth, including Fernandez, Ley, and Steel (2001), Sala-i Martin, Doppelhofer, and Miller (2004), Durlauf, Kourtellos, and Tan (2008), and Magnus, Powell, and Prufer (2010). We apply frequentist model averaging approaches as an alternative to Bayesian model averaging techniques to economic growth. We estimate the following cross-country growth regression

$$g_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + e_i \tag{8.1}$$

where g_i is average growth rate of GDP per capita between 1960 and 1996, $\mathbf{x}_i$ are the Solow variables from the neoclassical growth theory, and $\mathbf{z}_i$ are fundamental growth determinants such as geography, institutions, religion, and ethnic fractionalization from the new fundamental growth theory. Here, $\mathbf{x}_i$ are core regressors which appear in every submodels, while $\mathbf{z}_i$ are the auxiliary regressors which serve as controls of the neoclassical growth theory and may or may not be included in the submodels.

We follow Magnus, Powell, and Prufer (2010) and consider two model specifications to compare the neoclassical growth theory with the fundamental new growth theory. Model Setup A includes six core regressors and four auxiliary regressors. The six core regressors are the constant term (CONSTANT), the log of GDP per capita in 1960 (GDP60), the 1960-1985 equipment investment share of GDP (EQUIPINV), the primary school enrollment rate in 1960 (SCHOOL60), the life expectancy at age zero in 1960 (LIFE60), and the population growth rate between 1960 and 1990 (DPOP). The four auxiliary regressors are a rule of law index (LAW), a country's fraction of tropical area (TROPICS), an average index of ethnolinguistic fragmentation in a country (AVELF), and the

fraction of Confucian population (CONFUC), see Magnus, Powell, and Prufer (2010) for a detailed description of the data. Model Setup B contains only one core regressor, the constant term, and all other variables in Model Setup A are auxiliary regressors. The parameter of interest is the convergence term of the Solow growth model, that is, the coefficient of the log GDP per capita in 1960. The total number of observations is 74. We consider all possible submodels, that is, we have 16 submodels in Model Setup A and 512 submodels in Model Setup B.

We consider eight estimators: (1) the least-squares estimator for the full model (Full), (2) the averaging estimator with equal weights (Equal), (3) AIC model selection estimator (AIC), (4) BIC model selection estimator (BIC), (5) S-AIC model averaging estimator (S-AIC), (6) S-BIC model averaging estimator (S-BIC), (7) Jackknife Model Averaging estimator (JMA), and (8) Plug-In averaging estimator (Plug-In). The standard errors of data-driven model averaging estimators are calculated by (7.1).

The estimation results for Model Setup A and B are given in Table 5 and 6, respectively. We also repot the estimation results for weighted-average least-squares (WALS) estimator proposed by Magnus, Powell, and Prufer (2010) for comparison. The WALS estimator is a Bayesian model averaging technique which uses a Laplace distribution instead of the normal prior as the parameter prior. The results in Table 5 and 6 show that all coefficients have the same signs across different estimation methods. In model A, the coefficient estimate and standard error of GDP60 are similar between Plug-In, Full, Equal, and JMA estimators. Also, the 90% plug-in confidence interval for GDP60 is $(-0.0206, -0.0107)$, which is close to the naive confidence interval $(-0.0200, -0.0112)$.

In Model Setup B, the plug-in averaging estimate of GDP60 is quite close to the least-squares estimate from the full model and is higher in absolute value than other estimates. The 90% plug-in confidence interval for GDP60 is $(-0.0205, -0.0102)$, which is wider than the naive confidence interval $(-0.0183, -0.0124)$. The equal-weight averaging estimator has the smallest coefficient estimate and standard error of GDP60 because only half of submodels contains the regressor GDP60. The important finding from our results is that the plug-in averaging estimator has the smaller standard error of GDP60 than other estimators, except for the averaging estimator with equal weights.

It is also instructive to contrast the results of the Plug-In and WALS estimators. In Model Setup A, the estimation results are similar between Plug-In and WALS. In Model Setup B, the estimated coefficient of GDP60 is higher in absolute value for Plug-In than for WALS, while the estimated standard error of GDP60 is much smaller for Plug-In than for WALS. Therefore, the convergence speed of the growth model implied by our result is higher than that found by Magnus, Powell, and Prufer (2010). Comparing the results between Model Setup A and Model Setup B, we find that the plug-in averaging estimator chooses different fundamental growth determinants in different model specifications. Therefore, our results support the findings of Durlauf, Kourtellos, and Tan (2008) and Magnus, Powell, and Prufer (2010) that the fundamental variables are not robustly correlated with growth.

Table 7 and 8 report the weights placed on each submodel, and the regressor sets for each submodel are described in Table 9 and 10. We only report the results of AIC, BIC, JMA, and Plug-In estimators, since both S-AIC and S-BIC weights are spread out across all submodels. From

Table 7-10 we can see that AIC chooses a larger model than BIC in both model specifications, which is consistent with the previous literature. One interesting observation is that JMA and Plug-In choose completely different submodels in Model Setup A and B. The submodels chosen by JMA cover all entire regressor set, while Plug-In excludes the regressors LAW and TROPICS in Model Setup A and the regressors EQUIPINV, SCHOOL60, DPOP, and CONFUC in Model Setup B. Note that Plug-In puts 30% weight on the smallest submodel in Model Setup B. This particular model choice can explain the relatively small standard error of GDP60 of the plug-in estimate.

9 Conclusion

In this paper we study the frequentist model averaging estimator for heteroskedastic regressions in a local asymptotic framework. We characterize the optimal weights of the model averaging estimator and propose a plug-in estimator to estimate the infeasible optimal fixed weights. We derive the asymptotic distribution of the plug-in averaging estimator and suggest a plug-in method to construct the confidence interval. The simulation results show that the plug-in averaging estimator has much lower expected squared error than other selection and averaging methods. Also, the plug-in averaging estimator achieves the minimax risk and minimax regret in all simulations. We apply the plug-in averaging estimator to cross-country growth regressions. Our estimator has the smaller variance of the log GDP per capita in 1960, though our regression coefficient of the log GDP per capita in 1960 is close to those of other estimators. Our results also find little evidence of the new fundamental growth theory.

Appendix

A Proofs

Proof of Lemma 1: We first show the asymptotic distribution of the least-squares estimator for the full model. By Assumption 2 and the application of the continuous mapping theorem, it follows that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\left(\frac{1}{n}\mathbf{H}'\mathbf{H}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{H}'\mathbf{e}\right)\xrightarrow{d}\mathbf{Q}^{-1}\mathbf{R}\sim\mathbf{N}(\mathbf{0},\mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}).$$

We next show the asymptotic distribution of the least-squares estimator for each submodel. Define the extended selection matrix $\mathbf{S}_m$ as

$$\mathbf{S}_m=\begin{pmatrix}\mathbf{I}_k & \mathbf{0}_{k\times\ell_m}\\ \mathbf{0}_{\ell\times k} & \boldsymbol{\Pi}'_m\end{pmatrix}.$$

Then we have $\mathbf{H}_m=(\mathbf{X},\mathbf{Z}\boldsymbol{\Pi}'_m)=\mathbf{H}\mathbf{S}_m$ and $\boldsymbol{\Omega}_m=\mathbf{S}'_m\boldsymbol{\Omega}\mathbf{S}_m$. By some algebra, it follows that

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_m&=(\mathbf{H}'_m\mathbf{H}_m)^{-1}\mathbf{H}'_m\mathbf{y}\\&=\left(\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\left(\mathbf{H}'_m\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\boldsymbol{\gamma}+\mathbf{Z}(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m)\boldsymbol{\gamma}+\mathbf{e}\right)\right)\\&=\left(\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\mathbf{H}'_m\mathbf{H}_m\boldsymbol{\theta}_m+\left(\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\mathbf{H}'_m\mathbf{Z}\left(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+\left(\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\mathbf{H}'_m\mathbf{e}\\&=\boldsymbol{\theta}_m+\left(\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\mathbf{H}'_m\mathbf{Z}\left(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+\left(\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\mathbf{S}'_m\mathbf{H}'\mathbf{e}.\end{aligned}$$

Therefore, by Assumptions 1-2 and the application of the continuous mapping theorem, we have

$$\begin{aligned}\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_m-\boldsymbol{\theta}_m\right)&=\left(\frac{1}{n}\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\left(\frac{1}{n}\mathbf{H}'_m\mathbf{Z}\right)\left(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\sqrt{n}\boldsymbol{\gamma}+\left(\frac{1}{n}\mathbf{H}'_m\mathbf{H}_m\right)^{-1}\mathbf{S}'_m\left(\frac{1}{\sqrt{n}}\mathbf{H}'\mathbf{e}\right)\\&\xrightarrow{d}\mathbf{Q}_m^{-1}\begin{pmatrix}\mathbf{Q}_{\mathbf{xz}}\\ \boldsymbol{\Pi}_m\mathbf{Q}_{\mathbf{zz}}\end{pmatrix}\left(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{\delta}+\mathbf{Q}_m^{-1}\mathbf{S}'_m\mathbf{R}\\&=\mathbf{A}_m\boldsymbol{\delta}+\mathbf{B}_m\mathbf{R}\sim\mathbf{N}\left(\mathbf{A}_m\boldsymbol{\delta},\ \mathbf{Q}_m^{-1}\boldsymbol{\Omega}_m\mathbf{Q}_m^{-1}\right)\end{aligned}$$

where

$$\mathbf{A}_m=\mathbf{Q}_m^{-1}\begin{pmatrix}\mathbf{Q}_{\mathbf{xz}}\\ \boldsymbol{\Pi}_m\mathbf{Q}_{\mathbf{zz}}\end{pmatrix}\left(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\ \text{and}\ \mathbf{B}_m=\mathbf{Q}_m^{-1}\mathbf{S}'_m.$$

This completes the proof. ■

Proof of Lemma 2: Define $\boldsymbol{\gamma}_{m^c}=\{\boldsymbol{\gamma}:\gamma_j\notin\boldsymbol{\gamma}_m,\ for\ j=1,...,\ell\}$. That is, $\boldsymbol{\gamma}_{m^c}$ is the set of parameters γ_j which are not included in submodel m. Hence, we can write $\mu(\boldsymbol{\theta})$ as $\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\boldsymbol{\gamma}_{m^c})$. Also, $\mu(\boldsymbol{\theta}_m)=\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\mathbf{0})$.

Note that $\boldsymbol{\gamma}=O(n^{-1/2})$ by Assumption 1. Then by a standard Taylor series expansion of $\mu(\boldsymbol{\theta})$ about $\boldsymbol{\gamma}_{m^c}=\mathbf{0}$, it follows that

$$\begin{aligned}\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\boldsymbol{\gamma}_{m^c})&=\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\mathbf{0})+\mathbf{D}'_{\boldsymbol{\gamma}_{m^c}}\boldsymbol{\gamma}_{m^c}+O(n^{-1})\\&=\mu(\boldsymbol{\beta},\boldsymbol{\gamma}_m,\mathbf{0})+\mathbf{D}'_{\boldsymbol{\gamma}}\left(\mathbf{I}_\ell-\boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{\gamma}+O(n^{-1}).\end{aligned}$$

That is, $\mu(\boldsymbol{\theta}) - \mu(\boldsymbol{\theta}_m) = \mathbf{D}'_{\boldsymbol{\gamma}} \left(\mathbf{I}_\ell - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \boldsymbol{\gamma} + O(n^{-1})$. Thus, by Assumptions 1-2 and the application of the delta method, we have

$$\begin{aligned}
\sqrt{n}\left(\mu(\tilde{\boldsymbol{\theta}}_m) - \mu(\boldsymbol{\theta})\right) &= \sqrt{n}\left(\mu(\tilde{\boldsymbol{\theta}}_m) - \mu(\boldsymbol{\theta}_m)\right) - \sqrt{n}\Big(\mu(\boldsymbol{\theta}) - \mu(\boldsymbol{\theta}_m)\Big) \\
&\xrightarrow{d} \mathbf{D}'_{\boldsymbol{\theta}_m} \left(\mathbf{A}_m \boldsymbol{\delta} + \mathbf{B}_m \mathbf{R}\right) - \mathbf{D}'_{\boldsymbol{\gamma}} \left(\mathbf{I}_\ell - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \boldsymbol{\delta} \\
&= \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{A}_m \boldsymbol{\delta} - \mathbf{D}'_{\boldsymbol{\gamma}} \left(\mathbf{I}_\ell - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \boldsymbol{\delta} + \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{B}_m \mathbf{R} \\
&= \left(\mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \begin{pmatrix} \mathbf{Q}_{\mathbf{xz}} \\ \boldsymbol{\Pi}_m \mathbf{Q}_{\mathbf{zz}} \end{pmatrix} - \mathbf{D}'_{\boldsymbol{\gamma}}\right) \left(\mathbf{I}_\ell - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \boldsymbol{\delta} + \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \mathbf{S}'_m \mathbf{R} \\
&= \mathbf{a}'_m \boldsymbol{\delta} + \mathbf{b}'_m \mathbf{R} \equiv \Lambda_m \sim \mathbf{N}\left(\mathbf{a}'_m \boldsymbol{\delta},\ \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_m \mathbf{Q}_m^{-1} \mathbf{D}_{\boldsymbol{\theta}_m}\right),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{a}_m &= \left(\mathbf{I}_\ell - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \left(\begin{pmatrix} \mathbf{Q}_{\mathbf{zx}} \\ \mathbf{Q}_{\mathbf{zz}} \boldsymbol{\Pi}'_m \end{pmatrix} \mathbf{Q}_m^{-1} \mathbf{D}_{\boldsymbol{\theta}_m} - \mathbf{D}_{\boldsymbol{\gamma}}\right), \\
\mathbf{b}_m &= \mathbf{S}_m \mathbf{Q}_m^{-1} \mathbf{D}_{\boldsymbol{\theta}_m}.
\end{aligned}$$

This completes the proof. ■

Proof of Theorem 1: From Lemma 2, there is joint convergence in distribution of all $\sqrt{n}\left(\mu(\tilde{\boldsymbol{\theta}}_m) - \mu(\boldsymbol{\theta})\right)$ to Λ_m since all of Λ_m can be expressed in terms of $\mathbf{R}$. Since the weights are non-random, it follows that

$$\sqrt{n}\left(\bar{\mu}(\mathbf{w}) - \mu\right) = \sum_{m=1}^{M} w_m \sqrt{n}\left(\tilde{\mu}_m - \mu\right) \xrightarrow{d} \sum_{m=1}^{M} w_m \Lambda_m \equiv \Lambda.$$

Therefore, the asymptotic distribution of the averaging estimator is a weighted average of the normal distributions which is also a normal distribution.

By Lemma 2 and standard algebra, we can show the mean of Λ as

$$\mathrm{E}\left(\sum_{m=1}^{M} w_m \Lambda_m\right) = \sum_{m=1}^{M} w_m \mathrm{E}\left(\Lambda_m\right) = \sum_{m=1}^{M} w_m \mathbf{a}'_m \boldsymbol{\delta} = \mathbf{a}' \boldsymbol{\delta}, \text{ and } \mathbf{a} = \sum_{m=1}^{M} w_m \mathbf{a}_m.$$

Next we want to show the variance of Λ. For any two submodels, we have

$$\begin{aligned}
Cov(\Lambda_m, \Lambda_p) &= \mathrm{E}\left(\left(\mathbf{a}'_m \boldsymbol{\delta} + \mathbf{b}'_m \mathbf{R} - \mathrm{E}(\mathbf{a}'_m \boldsymbol{\delta} + \mathbf{b}'_m \mathbf{R})\right)\left(\mathbf{a}'_p \boldsymbol{\delta} + \mathbf{b}'_p \mathbf{R} - \mathrm{E}(\mathbf{a}'_p \boldsymbol{\delta} + \mathbf{b}'_p \mathbf{R})\right)\right) \\
&= \mathrm{E}\left(\mathbf{b}'_m \mathbf{R} \mathbf{b}'_p \mathbf{R}\right) \\
&= \mathbf{b}'_m \mathrm{E}\left(\mathbf{R}\mathbf{R}'\right) \mathbf{b}_p \\
&= \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \mathbf{S}'_m \boldsymbol{\Omega} \mathbf{S}_p \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p} \\
&= \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_{m,p} \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p}
\end{aligned}$$

with

$$\boldsymbol{\Omega}_{m,p} = \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\Omega}_{\mathbf{xz}} \boldsymbol{\Pi}'_p \\ \boldsymbol{\Pi}_m \boldsymbol{\Omega}_{\mathbf{zx}} & \boldsymbol{\Pi}_m \boldsymbol{\Omega}_{\mathbf{zz}} \boldsymbol{\Pi}'_p \end{pmatrix}$$

where the second equality holds by the fact that $\mathbf{a}_m$, $\mathbf{b}_m$, and $\boldsymbol{\delta}$ are constant vectors and $\mathbf{R} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$. Therefore, variance of Λ is

$$\begin{aligned} V &= var\left(\sum_{m=1}^{M} w_m \Lambda_m\right) \\ &= \sum_{m=1}^{M} w_m^2 Var(\Lambda_m) + 2\sum_{m<p} w_m w_p Cov(\Lambda_m, \Lambda_p) \\ &= \sum_{m=1}^{M} w_m^2 \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_m \mathbf{Q}_m^{-1} \mathbf{D}_{\boldsymbol{\theta}_m} + 2\sum_{m<p} w_m w_p \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_{m,p} \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p}. \end{aligned}$$

This completes the proof. ■

Proof of Theorem 2: We first show $\hat{\mathbf{D}}_{\boldsymbol{\theta}_m}$, $\hat{\mathbf{Q}}_m$, $\hat{\boldsymbol{\Omega}}_{m,p}$, and $\hat{\mathbf{a}}_m$ are consistent estimators for $\mathbf{D}_{\boldsymbol{\theta}_m}$, $\mathbf{Q}_m$, $\boldsymbol{\Omega}_{m,p}$, and $\mathbf{a}_m$. By Lemma 1, we have $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, which also implies that $\partial\mu(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta} = \hat{\mathbf{D}}_{\boldsymbol{\theta}} \xrightarrow{p} \mathbf{D}_{\boldsymbol{\theta}}$. By Assumption 2 and 3 and the fact that the selection matrix is non-random, we have $\hat{\mathbf{D}}_{\boldsymbol{\theta}_m} \xrightarrow{p} \mathbf{D}_{\boldsymbol{\theta}_m}$, $\hat{\mathbf{Q}}_m \xrightarrow{p} \mathbf{Q}_m$, and $\hat{\boldsymbol{\Omega}}_{m,p} \xrightarrow{p} \boldsymbol{\Omega}_{m,p}$. By Assumption 2 and the application of the continuous mapping theorem, it follows that $\hat{\mathbf{a}}_m \xrightarrow{p} \mathbf{a}_m$.

We next show the limiting distribution of $\hat{\zeta}_{m,p}$. By Assumption 2 and 3 and the application of the continuous mapping theorem, it follows that $\hat{\mathbf{D}}'_{\boldsymbol{\theta}_m} \hat{\mathbf{Q}}_m^{-1} \hat{\boldsymbol{\Omega}}_{m,p} \hat{\mathbf{Q}}_p^{-1} \hat{\mathbf{D}}_{\boldsymbol{\theta}_p} \xrightarrow{p} \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_{m,p} \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p}$. Recall that $\hat{\boldsymbol{\delta}} \xrightarrow{d} \mathbf{R}_{\boldsymbol{\delta}} = \boldsymbol{\delta} + \boldsymbol{\Pi}_\ell \mathbf{Q}^{-1}\mathbf{R}$. Then by the application of Slutsky's theorem, we have

$$\begin{aligned} \hat{\zeta}_{m,p} &= \hat{\boldsymbol{\delta}}' \hat{\mathbf{a}}_m \hat{\mathbf{a}}'_p \hat{\boldsymbol{\delta}} + \hat{\mathbf{D}}'_{\boldsymbol{\theta}_m} \hat{\mathbf{Q}}_m^{-1} \hat{\boldsymbol{\Omega}}_{m,p} \hat{\mathbf{Q}}_p^{-1} \hat{\mathbf{D}}_{\boldsymbol{\theta}_p} \\ &\xrightarrow{d} \mathbf{R}'_{\boldsymbol{\delta}} \mathbf{a}_m \mathbf{a}'_p \mathbf{R}_{\boldsymbol{\delta}} + \mathbf{D}'_{\boldsymbol{\theta}_m} \mathbf{Q}_m^{-1} \boldsymbol{\Omega}_{m,p} \mathbf{Q}_p^{-1} \mathbf{D}_{\boldsymbol{\theta}_p} = \zeta^*_{m,p}. \end{aligned}$$

Since all of $\zeta^*_{m,p}$ can be expressed in terms of the normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\hat{\zeta}_{m,p}$ to $\zeta^*_{m,p}$. Hence, it follows that $\mathbf{w}'\hat{\boldsymbol{\zeta}}\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$.

We now show the limiting distribution of $\hat{\mathbf{w}}_{pia}$. Note that $\mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$ is a convex minimization problem since $\mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$ is quadratic and $\boldsymbol{\zeta}^*$ is positive definite. Hence, the limiting process $\mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$ is continuous in $\mathbf{w}$ and has a unique minimum. Also note that $\hat{\mathbf{w}}_{pia} = O_p(1)$. By Theorem 3.2.2 of Van der Vaart and Wellner (1996) or Theorem 2.7 of Kim and Pollard (1990), the minimizer $\hat{\mathbf{w}}_{pia}$ converges in distribution to the minimizer of $\mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$, which is $\mathbf{w}^*_{pia}$.

Finally, we show the asymptotic distribution of the plug-in averaging estimator. Since both Λ_m and $w^*_{pia,m}$ can be expressed in terms of the same normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\tilde{\mu}_m$ and $\hat{w}_{pia,m}$. By Lemma 2, (4.2), and (4.9), it follows that

$$\sqrt{n}\big(\bar{\mu}(\hat{\mathbf{w}}_{pia}) - \mu\big) = \sum_{m=1}^{M} \hat{w}_{pia,m}\sqrt{n}\left(\tilde{\mu}_m - \mu\right) \xrightarrow{d} \sum_{m=1}^{M} w^*_{pia,m}\Lambda_m.$$

This completes the proof. ■

Proof of Theorem 3: By (5.2) and (5.3), it follows that $\hat{w}_{saic,m} \xrightarrow{d} w^*_{saic,m}$. Also, there is joint convergence in distribution of all $\hat{w}_{wsaic,m}$ and $\tilde{\mu}_m$. Thus, the limiting distribution of the S-AIC model averaging estimator follows from (5.1) and (5.3). This completes the proof. ■

Proof of Theorem 4: Define $h_i = \mathbf{h}_i'(\mathbf{H}'\mathbf{H})^{-1}\mathbf{h}_i$. Notice that $\hat{e}_{-i} = \hat{e}_i(1-h_i)^{-1} \approx \hat{e}_i(1+h_i)$ where $\hat{e}_i$ is the least-squares residuals and $\hat{e}_{-i}$ is the leave-one-out least-squares residual from the full model. For the submodel m, we have $\mathbf{h}_{m,i} = (\mathbf{x}_i', \mathbf{z}_i'\mathbf{\Pi}_m)' = (\mathbf{x}_i', \mathbf{z}_{mi}')'$, $h_{m,i} = \mathbf{h}_{m,i}'(\mathbf{H}_m'\mathbf{H}_m)^{-1}\mathbf{h}_{m,i}$, and $\tilde{e}_{m,-i} \approx \tilde{e}_{m,i}(1+h_{m,i})$.

Then it follows that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,-i}\tilde{e}_{p,-i} &\approx \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i} + \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i}(h_{m,i}+h_{p,i}) + \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i}h_{m,i}h_{p,i} \\
&= \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i} + \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i}\mathbf{h}_{m,i}'\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\mathbf{h}_{m,i} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i}\mathbf{h}_{p,i}'\left(\mathbf{H}_p'\mathbf{H}_p\right)^{-1}\mathbf{h}_{p,i} + o(1) \\
&= \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i} + \frac{1}{n}tr\left(\left(\mathbf{H}_m'\mathbf{H}_m\right)^{-1}\sum_{i=1}^{n}\mathbf{h}_{m,i}\mathbf{h}_{m,i}'\tilde{e}_{m,i}\tilde{e}_{p,i}\right) \\
&\quad + \frac{1}{n}tr\left(\left(\mathbf{H}_p'\mathbf{H}_p\right)^{-1}\sum_{i=1}^{n}\mathbf{h}_{p,i}\mathbf{h}_{p,i}'\tilde{e}_{m,i}\tilde{e}_{p,i}\right) + o(1) \\
&= \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_{m,i}\tilde{e}_{p,i} + \frac{1}{n}tr\left(\hat{\mathbf{Q}}_m^{-1}\tilde{\mathbf{\Omega}}_{m,m,p}\right) + \frac{1}{n}tr\left(\hat{\mathbf{Q}}_p^{-1}\tilde{\mathbf{\Omega}}_{p,m,p}\right) + o(1),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mathbf{Q}}_m &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{x}_i \\ \mathbf{z}_{mi}\end{pmatrix}\begin{pmatrix}\mathbf{x}_i' & \mathbf{z}_{mi}'\end{pmatrix}, \\
\tilde{\mathbf{\Omega}}_{m,m,p} &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{x}_i \\ \mathbf{z}_{mi}\end{pmatrix}\begin{pmatrix}\mathbf{x}_i' & \mathbf{z}_{mi}'\end{pmatrix}\tilde{e}_{m,i}\tilde{e}_{p,i}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\xi_{m,p} &= \tilde{\mathbf{e}}_{m,-i}'\tilde{\mathbf{e}}_{p,-i} - \hat{\mathbf{e}}'\hat{\mathbf{e}} \\
&= \left(\tilde{\mathbf{e}}_m'\tilde{\mathbf{e}}_p - \hat{\mathbf{e}}'\hat{\mathbf{e}}\right) + tr\left(\hat{\mathbf{Q}}_m^{-1}\tilde{\mathbf{\Omega}}_{m,m,p}\right) + tr\left(\hat{\mathbf{Q}}_p^{-1}\tilde{\mathbf{\Omega}}_{p,m,p}\right) + o(1), \qquad \text{(A.1)}
\end{aligned}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{H}\hat{\boldsymbol{\theta}}$ and $\tilde{\mathbf{e}}_{\mathbf{m}} = \mathbf{y} - \mathbf{H}_m\tilde{\boldsymbol{\theta}}_m$.

First, we consider the first terms of (A.1). Since $\tilde{\mathbf{e}}_m'\hat{\mathbf{e}} = \hat{\mathbf{e}}'\hat{\mathbf{e}}$ and $\tilde{\mathbf{e}}_m - \hat{\mathbf{e}} = \mathbf{H}(\mathbf{S}_m\tilde{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}})$, we have

$$\begin{aligned}
\tilde{\mathbf{e}}_m'\tilde{\mathbf{e}}_p - \hat{\mathbf{e}}'\hat{\mathbf{e}} &= \left(\tilde{\mathbf{e}}_m - \hat{\mathbf{e}}\right)'\left(\tilde{\mathbf{e}}_p - \hat{\mathbf{e}}\right) \\
&= \sqrt{n}(\hat{\boldsymbol{\theta}} - \mathbf{S}_m\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})'\left(\frac{1}{n}\mathbf{H}'\mathbf{H}\right)\sqrt{n}(\hat{\boldsymbol{\theta}} - \mathbf{S}_p\tilde{\boldsymbol{\theta}}_p).
\end{aligned}$$

Define $\mathbf{\Pi}_\ell = (\mathbf{0}_{\ell\times k}, \mathbf{I}_\ell)$. Then from Lemma 1 it follows that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}} - \mathbf{S}_m\tilde{\boldsymbol{\theta}}_m) &= \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \mathbf{S}_m\sqrt{n}(\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m) + \sqrt{n}(\boldsymbol{\theta} - \mathbf{S}_m\boldsymbol{\theta}_m) \\
&\xrightarrow{d} \mathbf{Q}^{-1}\mathbf{R} - \mathbf{S}_m\left(\mathbf{Q}_m^{-1}\begin{pmatrix}\mathbf{Q_{xz}}\\ \mathbf{\Pi}_m\mathbf{Q_{zz}}\end{pmatrix}\left(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m\right)\boldsymbol{\delta} + \mathbf{Q}_m^{-1}\mathbf{S}_m'\mathbf{R}\right) \\
&\quad + \begin{pmatrix}\mathbf{0}_{k\times 1}\\ \left(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m\right)\boldsymbol{\delta}\end{pmatrix} \\
&= \left(\mathbf{Q}^{-1} - \mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\right)\mathbf{R} + \left(\mathbf{\Pi}_\ell' - \mathbf{S}_m\mathbf{Q}_m^{-1}\begin{pmatrix}\mathbf{Q_{xz}}\\ \mathbf{\Pi}_m\mathbf{Q_{zz}}\end{pmatrix}\right)\left(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m\right)\boldsymbol{\delta} \\
&= \ddot{\mathbf{A}}_m\boldsymbol{\delta} + \ddot{\mathbf{B}}_m\mathbf{R} \equiv \ddot{\mathbf{R}}_m
\end{aligned}$$

where

$$\ddot{\mathbf{A}}_m = \left(\mathbf{\Pi}_\ell' - \mathbf{S}_m\mathbf{Q}_m^{-1}\begin{pmatrix}\mathbf{Q_{xz}}\\ \mathbf{\Pi}_m\mathbf{Q_{zz}}\end{pmatrix}\right)\left(\mathbf{I}_\ell - \mathbf{\Pi}_m'\mathbf{\Pi}_m\right), \text{ and } \ddot{\mathbf{B}}_m = \left(\mathbf{Q}^{-1} - \mathbf{S}_m\mathbf{Q}_m^{-1}\mathbf{S}_m'\right).$$

Therefore, it follows that

$$\tilde{\mathbf{e}}_m'\tilde{\mathbf{e}}_p - \hat{\mathbf{e}}'\hat{\mathbf{e}} \xrightarrow{d} \ddot{\mathbf{R}}_m'\mathbf{Q}\ddot{\mathbf{R}}_p. \tag{A.2}$$

Next, consider the second and third terms of (A.1). From Lemma 3 and the application of the continuous mapping theorem, it follows that

$$tr(\hat{\mathbf{Q}}_m^{-1}\tilde{\mathbf{\Omega}}_{m,m,p}) \xrightarrow{p} tr(\mathbf{Q}_m^{-1}\mathbf{\Omega}_m), \tag{A.3}$$

$$tr(\hat{\mathbf{Q}}_p^{-1}\tilde{\mathbf{\Omega}}_{p,m,p}) \xrightarrow{p} tr(\mathbf{Q}_p^{-1}\mathbf{\Omega}_p), \tag{A.4}$$

Equation (5.9) then follows from (A.2), (A.3), and (A.4). Since all of $\xi_{m,p}^*$ can be expressed in terms of the normal random vector $\mathbf{R}$, there is joint convergence in distribution of all $\xi_{m,p}$ to $\xi_{m,p}^*$. Hence, it follows that $\mathbf{w}'\boldsymbol{\xi}_n\mathbf{w} \xrightarrow{d} \mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}$.

Finally, we show the limiting distribution of $\hat{\mathbf{w}}_{jma}$ and $\bar{\mu}(\hat{\mathbf{w}}_{jma})$. First, the limiting process $\mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}$ is continuous in $\mathbf{w}$ and has a unique minimum since $\mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}$ is quadratic and $\boldsymbol{\xi}^*$ is positive definite. Second, $\hat{\mathbf{w}}_{jma} = O_p(1)$ by the fact that $\mathcal{H}_n$ is convex. Therefore, by Theorem 3.2.2 of Van der Vaart and Wellner (1996) or Theorem 2.7 of Kim and Pollard (1990), the minimizer $\hat{\mathbf{w}}_{jma}$ converges in distribution to the minimizer of $\mathbf{w}'\boldsymbol{\xi}^*\mathbf{w}$, which is $\mathbf{w}_{jma}^*$. Equation (5.11) then follows from the distribution result (5.10) and the fact that there is joint convergence in distribution of $\tilde{\mu}_m$ and $\hat{\mathbf{w}}_{jma}$. This completes the proof. ■

Lemma 3. *Let $\tilde{e}_{m,i} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_m - \mathbf{z}_{mi}'\hat{\boldsymbol{\gamma}}_m$ denote the OLS residuals from the submodels and*

$$\tilde{\mathbf{\Omega}}_{m,m,p} = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\mathbf{x}_i\\ \mathbf{z}_{mi}\end{pmatrix}\begin{pmatrix}\mathbf{x}_i' & \mathbf{z}_{mi}'\end{pmatrix}\tilde{e}_{m,i}\tilde{e}_{p,i}$$

for $m, p = 1, ..., M$. Suppose Assumptions 1 and 4 hold. As $n \to \infty$, we have

$$\tilde{\mathbf{\Omega}}_{m,m,p} \xrightarrow{p} \mathbf{\Omega}_m.$$

Proof of Lemma 3: Let $\|\cdot\|$ be the Euclidean norm. That is, for an $m \times n$ matrix $\mathbf{X}$, $\|\mathbf{X}\| = (\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2)^{1/2}$. Note that

$$\begin{aligned}
\tilde{e}_{m,i} &= y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}} - \mathbf{z}_{mi}'\tilde{\boldsymbol{\gamma}}_m \\
&= e_i - \mathbf{x}_i'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - (\mathbf{z}_{mi}'\tilde{\boldsymbol{\gamma}}_m - \mathbf{z}_i'\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\boldsymbol{\gamma}) + \mathbf{z}_i'(\mathbf{I}_\ell - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m)\boldsymbol{\gamma} \\
&= e_i - \Big(\mathbf{x}_i'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{z}_{mi}'(\tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m)\Big) + \mathbf{z}_{m^c i}'\boldsymbol{\gamma}_{m^c}
\end{aligned}$$

where $\mathbf{z}_{m^c i} = \{\mathbf{z}_i : z_{ji} \notin \mathbf{z}_{mi},\ for\ j = 1, ..., \ell\}$ and $\boldsymbol{\gamma}_{m^c} = \{\boldsymbol{\gamma} : \gamma_j \notin \boldsymbol{\gamma}_m,\ for\ j = 1, ..., \ell\}$.

Thus,

$$\begin{aligned}
\tilde{\boldsymbol{\Omega}}_{m,m,p} = &\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i'e_i^2 & \mathbf{x}_i\mathbf{z}_{mi}'e_i^2 \\ \mathbf{z}_{mi}\mathbf{x}_i'e_i^2 & \mathbf{z}_{mi}\mathbf{z}_{mi}'e_i^2 \end{pmatrix} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix}\begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix}'\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix}\begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{pmatrix}'\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p \end{pmatrix} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix}\mathbf{z}_{m^c i}'\boldsymbol{\gamma}_{m^c}\mathbf{z}_{p^c i}'\boldsymbol{\gamma}_{p^c} \\
&- \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix}\begin{pmatrix} \mathbf{x}_i e_i \\ \mathbf{z}_{mi} e_i \end{pmatrix}'\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \\
&- \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix}\begin{pmatrix} \mathbf{x}_i e_i \\ \mathbf{z}_{pi} e_i \end{pmatrix}'\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p \end{pmatrix} \\
&- \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix}\boldsymbol{\gamma}_{p^c}'\mathbf{z}_{p^c i}\begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix}'\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \\
&- \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix}\boldsymbol{\gamma}_{m^c}'\mathbf{z}_{m^c i}\begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{pmatrix}'\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p \end{pmatrix} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} e_i\mathbf{z}_{m^c i}'\boldsymbol{\gamma}_{m^c} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} e_i\mathbf{z}_{p^c i}'\boldsymbol{\gamma}_{p^c}
\end{aligned} \tag{A.5}$$

The strategy of the proof is to show that the first term of (A.5) converges in probability to $\boldsymbol{\Omega}_m$ and the remaining terms of (A.5) converge in probability to zero. First consider the first term of (A.5). The jl'th element of $\mathbf{x}_i\mathbf{x}_i'e_i^2$ is $x_{ji}x_{li}e_i^2$. By Assumption 4 and the application of Cauchy-Schwarz Inequality, it follows that

$$\mathrm{E}\left|x_{ji}x_{li}e_i^2\right| \leq \left(Ex_{ji}^2x_{li}^2\right)^{1/2}\left(Ee_i^4\right)^{1/2} \leq \left(Ex_{ji}^4\right)^{1/4}\left(Ex_{li}^4\right)^{1/4}\left(Ee_i^4\right)^{1/2} < \infty.$$

Similarly, we can show that the expectations of $|x_{ji}z_{m_{li}}e_i^2|$, $|z_{m_{ji}}x_{li}e_i^2|$, and $|z_{m_{ji}}z_{m_{li}}e_i^2|$ are finite. Then by weak law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{x}_i\mathbf{x}_i'e_i^2 & \mathbf{x}_i\mathbf{z}_{mi}'e_i^2 \\ \mathbf{z}_{mi}\mathbf{x}_i'e_i^2 & \mathbf{z}_{mi}\mathbf{z}_{mi}'e_i^2 \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \mathrm{E}\left(\mathbf{x}_i\mathbf{x}_i'e_i^2\right) & \mathrm{E}\left(\mathbf{x}_i\mathbf{z}_{mi}'e_i^2\right) \\ \mathrm{E}\left(\mathbf{z}_{mi}\mathbf{x}_i'e_i^2\right) & \mathrm{E}\left(\mathbf{z}_{mi}\mathbf{z}_{mi}'e_i^2\right) \end{pmatrix} = \boldsymbol{\Omega}_m.$$

Next consider the second term of (A.5). By the Triangle Inequality and Schwarz Inequality, it follows that

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix}' \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{pmatrix}' \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p \end{pmatrix} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \begin{matrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{matrix} \right\| \left| \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix}' \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right| \left| \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p \end{pmatrix}' \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{pmatrix} \right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^3 \left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{matrix} \right\| \left\| \begin{matrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{matrix} \right\| \left\| \begin{matrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_p - \boldsymbol{\gamma}_p \end{matrix} \right\|. \qquad \text{(A.6)}
\end{aligned}$$

Since from Lemma 1

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \longrightarrow \mathbf{0}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^3 \left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{matrix} \right\| \longrightarrow \mathrm{E}\left(\left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^3 \left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{pi} \end{matrix} \right\| \right) < \infty$$

it follows that (A.6) converges in probability to zero. This shows that the second term of (A.5) converges in probability to zero.

Next consider the third term of (A.5). By the Triangle Inequality and Schwarz Inequality, it follows that

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} \mathbf{z}_{m^ci}'\boldsymbol{\gamma}_{m^c}\mathbf{z}_{p^ci}'\boldsymbol{\gamma}_{p^c} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \begin{matrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{matrix} \right\| \left| \mathbf{z}_{m^ci}'\boldsymbol{\gamma}_{m^c} \right| \left| \mathbf{z}_{p^ci}'\boldsymbol{\gamma}_{p^c} \right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^2 \left\| \mathbf{z}_{m^ci} \right\| \left\| \mathbf{z}_{p^ci} \right\| \right) \left\| \boldsymbol{\gamma}_{m^c} \right\| \left\| \boldsymbol{\gamma}_{p^c} \right\|.
\end{aligned}$$

By the Cauchy-Schwarz Inequality, it follows that

$$\mathrm{E}\left(\left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^2 \left\| \mathbf{z}_{m^ci} \right\| \left\| \mathbf{z}_{p^ci} \right\| \right) \leq \mathrm{E}\left(\left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^4 \right)^{1/2} \mathrm{E}\left(\left\| \mathbf{z}_{m^ci} \right\|^2 \left\| \mathbf{z}_{p^ci} \right\|^2 \right)^{1/2} < \infty.$$

Then by Weak Law of Large Number,

$$\frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^2 \left\| \mathbf{z}_{m^ci} \right\| \left\| \mathbf{z}_{p^ci} \right\| \right) \longrightarrow \mathrm{E}\left(\left\| \begin{matrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{matrix} \right\|^2 \left\| \mathbf{z}_{m^ci} \right\| \left\| \mathbf{z}_{p^ci} \right\| \right) < \infty.$$

By Assumption 1, we have $\boldsymbol{\gamma}_{m^c} \to \mathbf{0}$ and $\boldsymbol{\gamma}_{p^c} \to \mathbf{0}$. Hence, the third term of (A.5) converges in probability to zero.

Next consider the fourth term of (A.5). By the Triangle Inequality and Schwarz Inequality, it follows that

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} \begin{pmatrix} \mathbf{x}_i e_i \\ \mathbf{z}_{mi} e_i \end{pmatrix}' \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} \right\| \left\| \begin{pmatrix} \mathbf{x}_i e_i \\ \mathbf{z}_{mi} e_i \end{pmatrix} \right\| \right) \left\| \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^3 |e_i| \right) \left\| \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \right\|. \qquad \text{(A.7)}
\end{aligned}$$

By Holder's Inequality, we have

$$\mathrm{E}\left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^3 |e_i| \right) \leq \mathrm{E}\left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^4 \right)^{3/4} \left(\mathrm{E}|e_i^4|\right)^{1/4} < \infty.$$

Then by Weak Law of Large Number,

$$\frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^3 |e_i| \right) \longrightarrow \mathrm{E}\left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^3 |e_i| \right) < \infty.$$

Therefore, (A.7) converges in probability to zero. This shows that the fourth term of (A.5) converges in probability to zero. Similarly, we can show the fifth term of (A.5) converges in probability to zero.

Next consider the sixth term of (A.5). By the Triangle Inequality and Schwarz Inequality, it follows that

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} \boldsymbol{\gamma}_{p^c}'\mathbf{z}_{pi} \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix}' \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^3 \|\mathbf{z}_{p^c i}\| \right) \|\boldsymbol{\gamma}_{p^c}\| \left\| \begin{pmatrix} \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_m \end{pmatrix} \right\| \\
&\longrightarrow 0.
\end{aligned}$$

Therefore, the sixth term of (A.5) converges in probability to zero. Similarly, it shows that the seventh term of (A.5) converges in probability to zero.

Next consider the eighth term of (A.5). By the Triangle Inequality and Schwarz Inequality, it follows that

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_i\mathbf{x}_i' & \mathbf{x}_i\mathbf{z}_{mi}' \\ \mathbf{z}_{mi}\mathbf{x}_i' & \mathbf{z}_{mi}\mathbf{z}_{mi}' \end{pmatrix} e_i \mathbf{z}_{m^c i}'\boldsymbol{\gamma}_{m^c} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(\left\| \begin{pmatrix} \mathbf{x}_i \\ \mathbf{z}_{mi} \end{pmatrix} \right\|^2 \|\mathbf{z}_{m^c i}\| \, |e_i| \right) \|\boldsymbol{\gamma}_{m^c}\| \\
&\longrightarrow 0.
\end{aligned}$$

It follows that the eighth and ninth terms of (A.5) converge in probability to zero. This completes the proof. ■

Proof of Corollary 1: From Theorem 1, we can express the AMSE of the averaging estimator for the two-model case as $\text{AMSE}(\bar{\mu}(w)) = w^2\zeta_{1,1} + (1-w)^2\zeta_{2,2} + 2w(1-w)\zeta_{1,2}$. The first-order condition for the minimization is $0 = 2w(\zeta_{1,1} + \zeta_{2,2} - 2\zeta_{1,2}) - 2(\zeta_{2,2} - \zeta_{1,2})$, whose solution is $w^o = (\zeta_{2,2} - \zeta_{1,2})/(\zeta_{1,1} + \zeta_{2,2} - 2\zeta_{1,2})$. If this value is greater than one, then the constrained minimizer is $w^o = 1$. If this value is negative, then the constrained minimizer is $w^o = 0$. This completes the proof. ■

Proof of Corollary 2: In Theorem 2, we show that $\hat{\mathbf{w}}_{pia} \xrightarrow{d} \mathbf{w}^*_{pia} = \underset{\mathbf{w}\in\mathcal{H}_n}{\text{argmin}}\, \mathbf{w}'\boldsymbol{\zeta}^*\mathbf{w}$. For $M = 2$, we have $w^*_{pia} = \underset{w\in\mathcal{H}_n}{\text{argmin}} \left(w^2\zeta^*_{1,1} + (1-w)^2\zeta^*_{2,2} + 2w(1-w)\zeta^*_{1,2}\right) = (\zeta^*_{2,2} - \zeta^*_{1,2})/(\zeta^*_{1,1} + \zeta^*_{2,2} - 2\zeta^*_{1,2})$. Therefore, $\text{AMSE}(\bar{\mu}(\hat{w}_{pia})) = \text{E}\big(w^{*2}_{pia}\zeta_{1,1} + \big(1 - w^*_{pia}\big)^2\zeta_{2,2} + 2w^*_{pia}\big(1 - w^*_{pia}\big)\zeta_{1,2}\big)$. The argument for w^*_{jma} is similar. This completes the proof. ■

Proof of Theorem 5: For any data-driven weights, we have $\hat{w}_m \xrightarrow{d} w^*_m$ where w^*_m is a function of the random vector $\mathbf{R}$. In Theorem 2, we show that $\hat{\mathbf{D}}'_{\boldsymbol{\theta}_m}\hat{\mathbf{Q}}^{-1}_m\hat{\boldsymbol{\Omega}}_{m,p}\hat{\mathbf{Q}}^{-1}_p\hat{\mathbf{D}}_{\boldsymbol{\theta}_p} \xrightarrow{p} \mathbf{D}'_{\boldsymbol{\theta}_m}\mathbf{Q}^{-1}_m\boldsymbol{\Omega}_{m,p}\mathbf{Q}^{-1}_p\mathbf{D}_{\boldsymbol{\theta}_p}$. Then by the application of Slutsky's theorem, we have $\hat{V} \xrightarrow{d} V^*$. From Theorems 2, 3, and 4, we show that $\sqrt{n}(\bar{\mu}(\hat{\mathbf{w}}) - \mu) \xrightarrow{d} \sum_{m=1}^M w^*_m\Lambda_m$ for some data-driven weights $\hat{\mathbf{w}}$. Therefore, there is joint convergence in distribution of $\hat{V}$ and $\sqrt{n}(\bar{\mu}(\hat{\mathbf{w}}) - \mu)$ since all of V^*, w_m, and Λ_m can be expressed in terms of $\mathbf{R}$. Finally, by the application of the continuous mapping theorem, it follows that $t_n(\mu) \xrightarrow{d} (V^*)^{-1/2}\sum_{m=1}^M w^*_m\Lambda_m$. This completes the proof. ■

B Figures

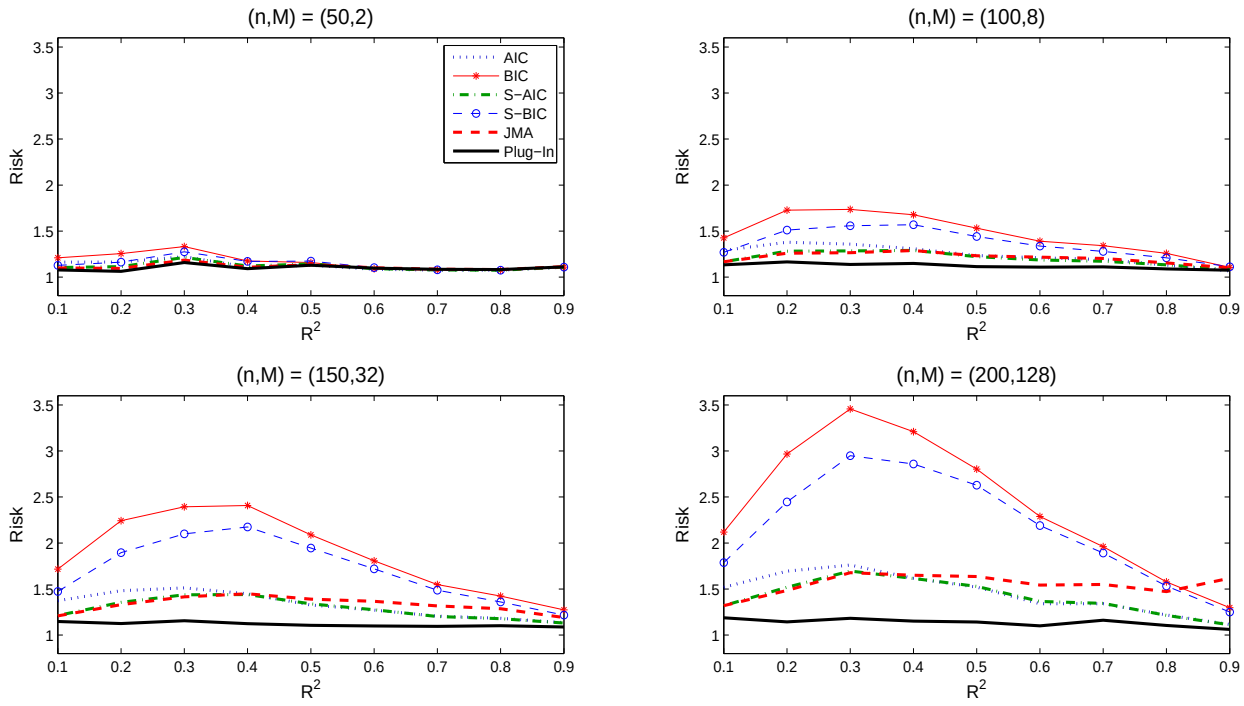

Figure 1: DGP_1, $\sigma_i^2 = 1$, $\rho_1 = 0.3$, $\rho_2 = 0.1$.

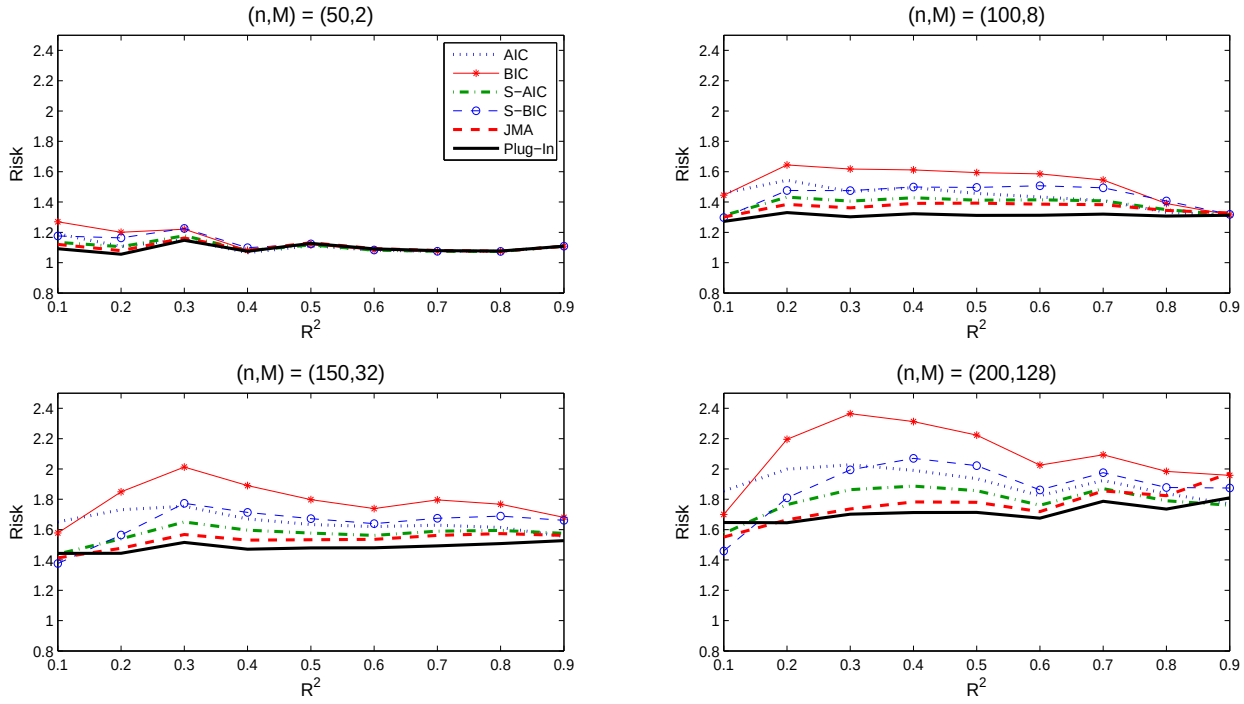

Figure 2: DGP_2, $\sigma_i^2 = 1$, $\rho_1 = 0.3$, $\rho_2 = 0.1$.

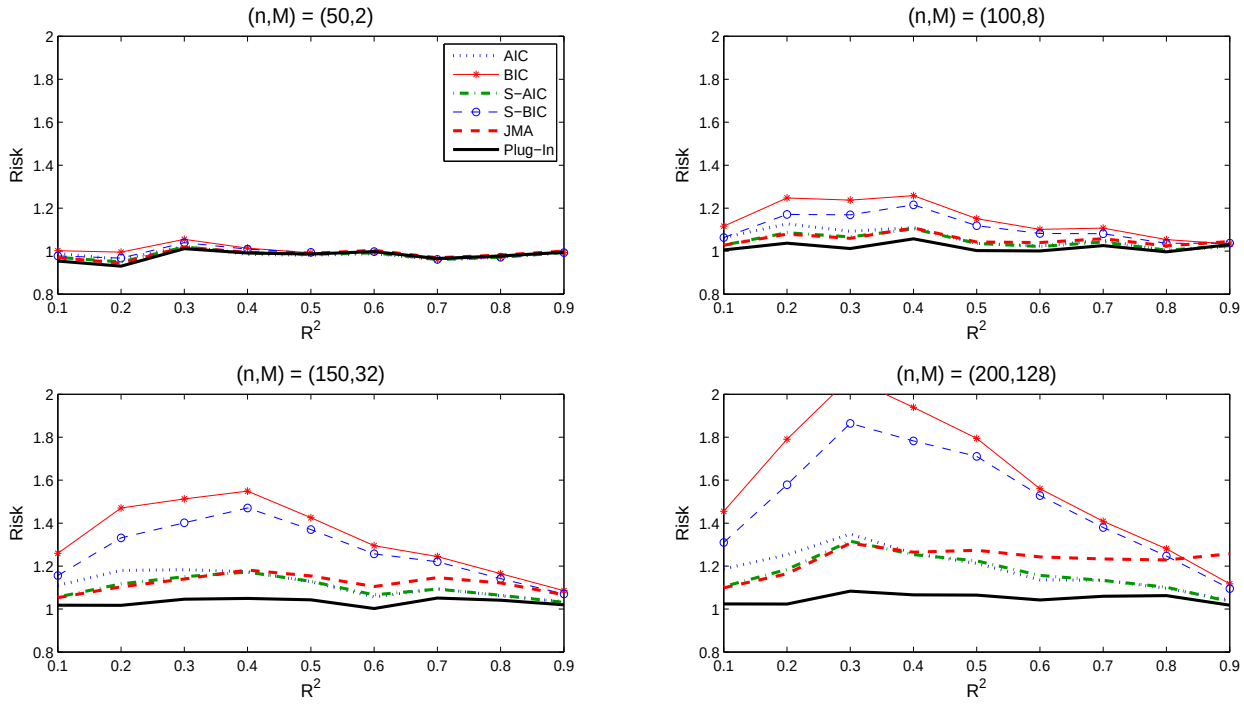

Figure 3: DGP_1, $\sigma_i^2 = x_{2i}^2$, $\rho_1 = 0.3$, $\rho_2 = 0.1$.

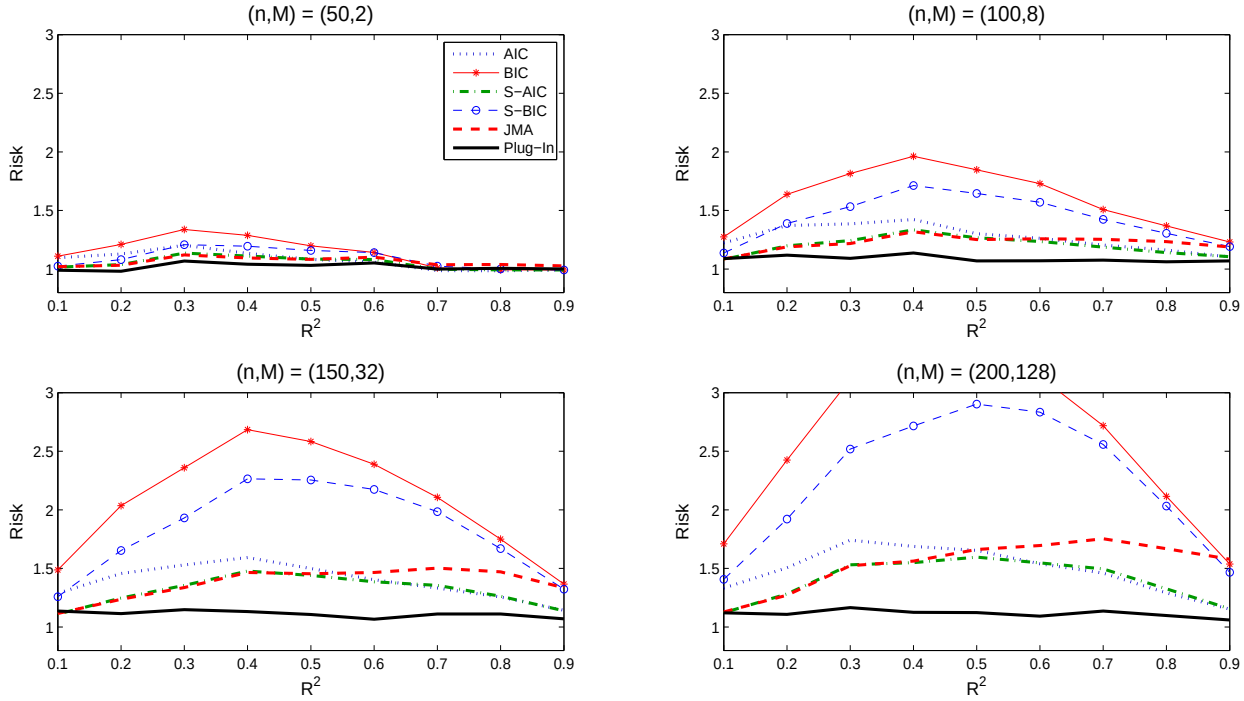

Figure 4: DGP_1, $\sigma_i^2 = x_{2i}^2$, $\rho_1 = 0.6$, $\rho_2 = 0.4$.

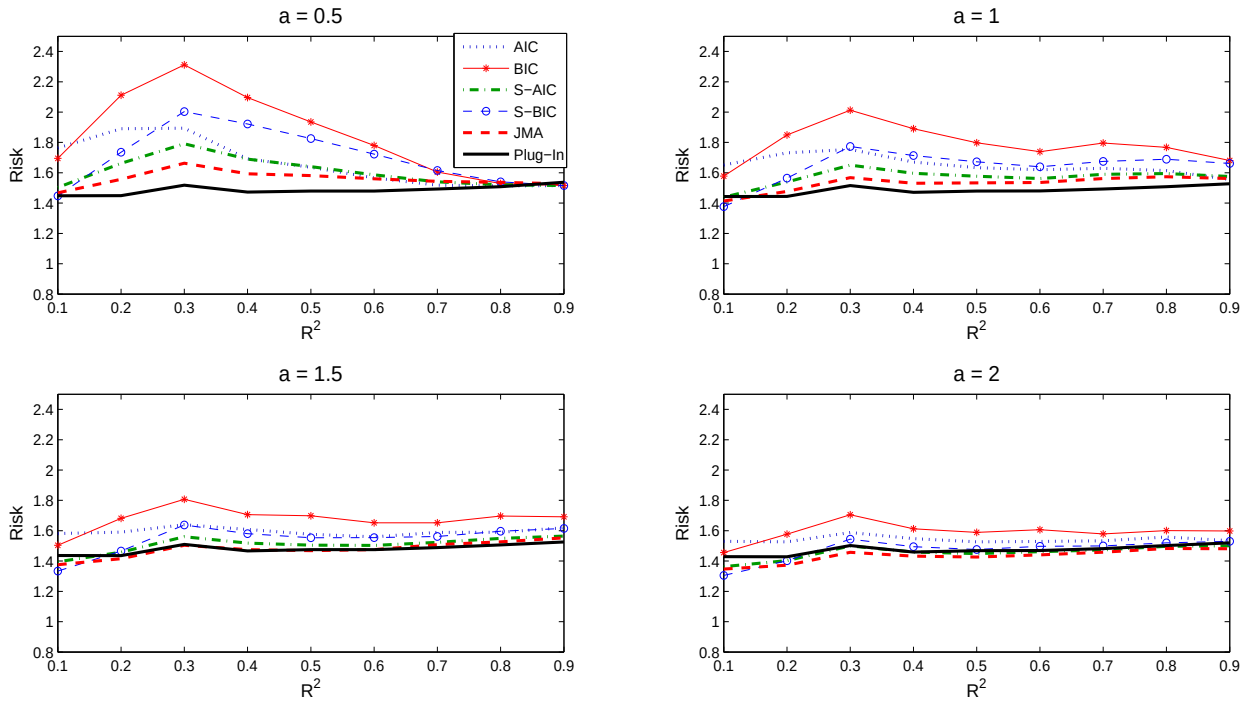

Figure 5: DGP_3, $\sigma_i^2 = 1$, $\rho_1 = 0.3$, $\rho_2 = 0.1$.

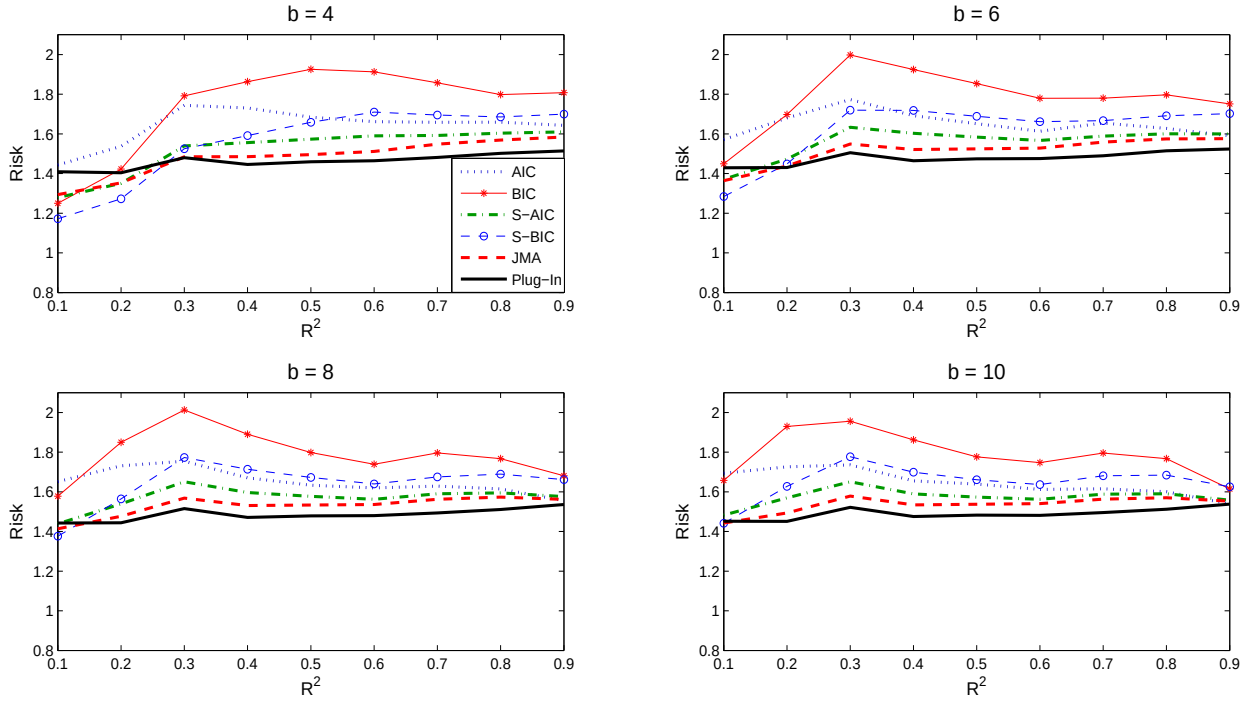

Figure 6: DGP_4, $\sigma_i^2 = 1$, $\rho_1 = 0.3$, $\rho_2 = 0.1$.

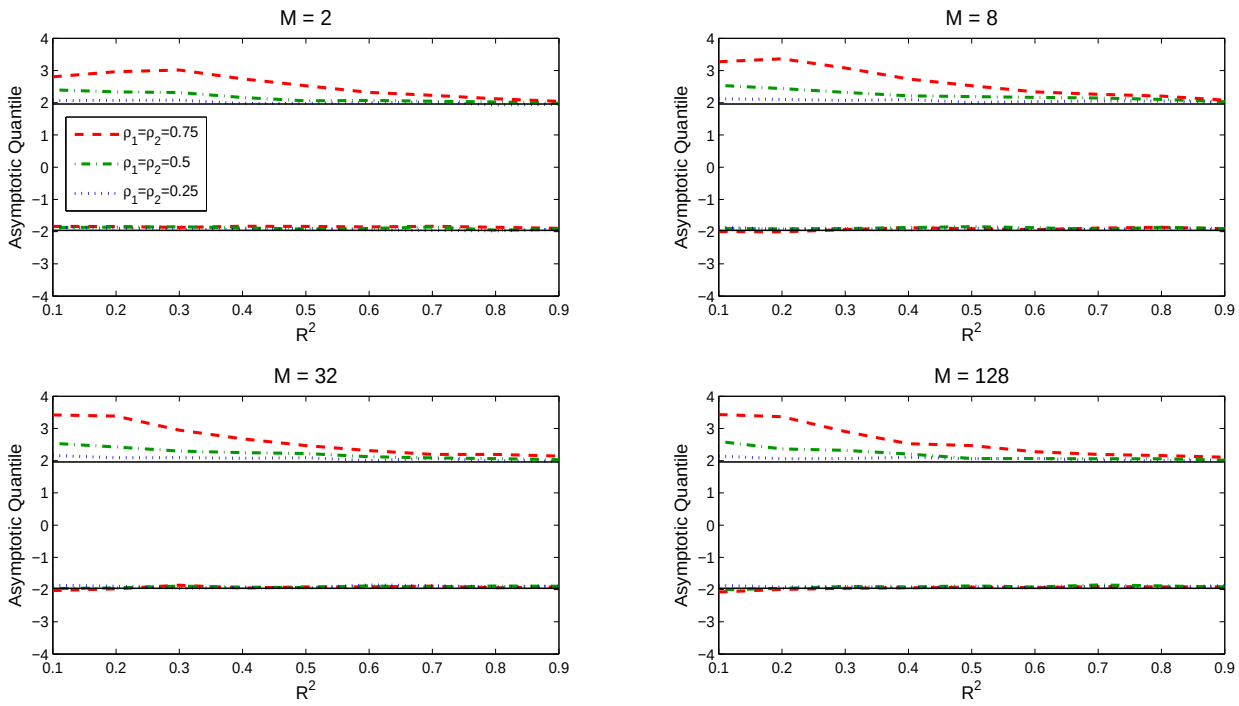

Figure 7: DGP_1, $\sigma_i^2 = 1$, $\alpha = 0.05$.

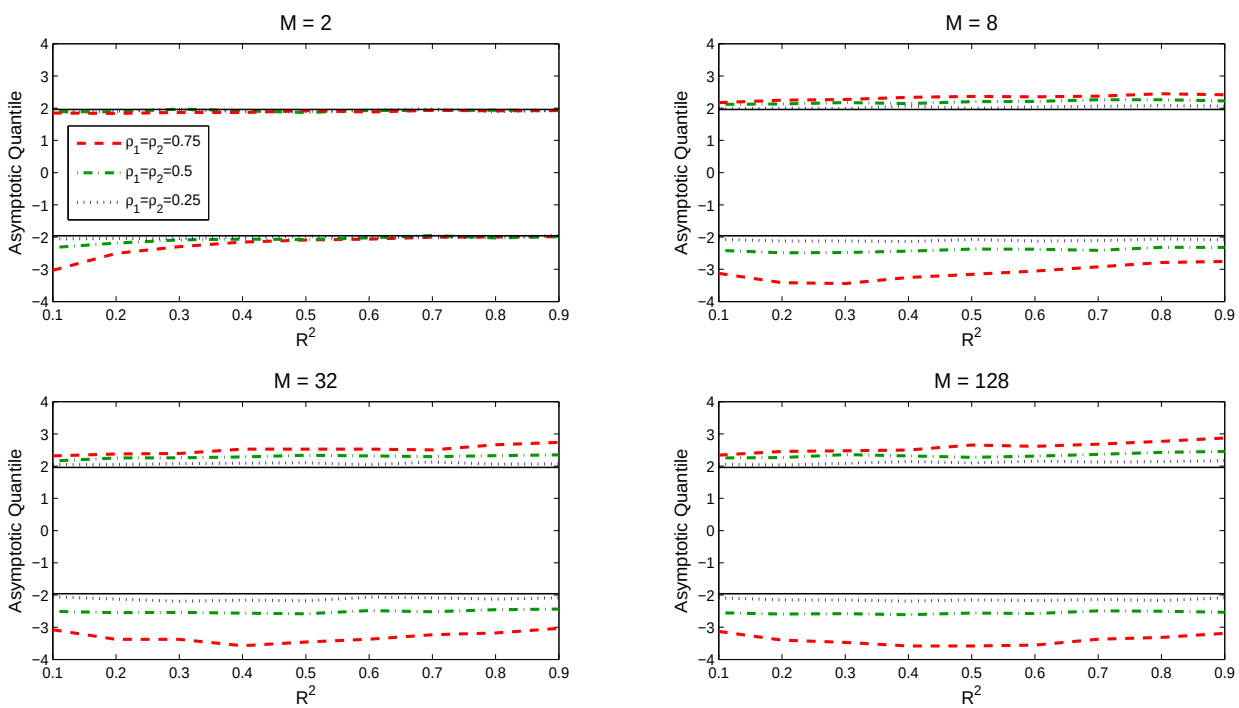

Figure 8: DGP_2, $\sigma_i^2 = 1$, $\alpha = 0.05$.

C Tables

Table 1: Maximum Risk

DGP	M	AIC	BIC	S-AIC	S-BIC	JMA	Plug-In	Opt
A. Homoskedastic error								
1	2	2.4029	2.9376	2.1730	2.4257	2.0733	1.8919	1.5570
	8	4.0453	6.3674	3.6617	5.2650	3.6230	2.8533	2.4557
	32	5.7872	11.1803	5.2191	9.3306	5.5725	3.5633	3.1532
	128	7.2794	16.7302	6.8562	14.0665	7.8083	4.3867	3.7290
2	2	2.2995	2.7909	2.0511	2.3132	1.9469	1.8717	1.5611
	8	3.5464	4.0076	3.1097	3.4690	2.9640	2.6806	1.2386
	32	4.3184	5.1901	3.8780	4.2957	3.6486	3.4532	1.1603
	128	5.1320	6.1632	4.6347	5.2107	5.1819	4.6161	1.1236
B. Heteroskedastic error								
1	2	4.1706	4.6361	3.9392	4.1876	3.8890	3.7151	3.5570
	8	6.0116	8.2923	5.6419	7.2394	5.5690	4.8047	4.4557
	32	7.7131	13.0020	7.1562	11.0792	7.5651	5.6508	5.1532
	128	9.1700	18.4394	8.7134	15.8386	9.8053	6.3532	5.7290
2	2	4.0088	4.4651	3.8047	4.0485	3.8256	3.6676	3.5611
	8	5.3354	5.8515	4.9670	5.2761	4.9241	4.5924	3.2386
	32	6.2868	7.2808	5.9291	6.3404	5.7661	5.4059	3.1603
	128	6.8859	8.0384	6.5151	7.0798	7.1793	6.4366	3.1236

Table 2: Maximum Regret

DGP	M	AIC	BIC	S-AIC	S-BIC	JMA	Plug-In
A. Homoskedastic error							
1	2	0.9369	1.4717	0.7071	0.9598	0.6073	0.4260
	8	1.8198	4.1419	1.4363	3.0297	1.3320	0.6384
	32	2.9441	8.3371	2.3216	6.4162	2.5958	0.8282
	128	3.9048	13.2733	3.3993	10.6096	4.2170	1.0019
2	2	0.8318	1.3232	0.5834	0.8454	0.4792	0.3373
	8	2.3289	2.7902	1.8791	2.2337	1.7335	1.4428
	32	3.1657	4.0373	2.7253	3.1429	2.4887	2.2929
	128	4.0135	5.0447	3.5144	4.0904	4.0582	3.4925
B. Heteroskedastic error							
1	2	0.7047	1.1702	0.4733	0.7217	0.4190	0.2365
	8	1.7861	4.0669	1.4165	3.0139	1.3435	0.5793
	32	2.8699	8.1588	2.3130	6.1647	2.5310	0.7052
	128	3.9034	12.9825	3.2565	10.3817	4.2141	0.8709
2	2	0.5411	0.9974	0.3370	0.5808	0.2967	0.1332
	8	2.1179	2.6340	1.7364	2.0408	1.6935	1.3538
	32	3.1340	4.1281	2.7764	3.1876	2.6063	2.2461
	128	3.7656	4.9199	3.3948	3.9594	4.0557	3.3129

Table 3: Maximum Risk and Regret

	DGP	AIC	BIC	S-AIC	S-BIC	JMA	Plug-In
Maximum Risk	3	1.9428	2.3719	1.8376	2.0556	1.7060	1.5789
	4	1.8158	2.0624	1.6916	1.8171	1.6250	1.5758
Maximum Regret	3	0.9168	1.3459	0.8116	1.0296	0.6800	0.5518
	4	0.7916	1.0376	0.6669	0.7924	0.5993	0.5500

Table 4: Coverage Probabilities of 90% and 95% Confidence Intervals for $\sigma_i^2 = 1$, and $\rho_1 = \rho_2 = 0.75$

			90%			95%		
M	DGP	R^2	Naive	BBA	Plug-In	Naive	BBA	Plug-In
2	1	0.1	0.7940	0.8540	0.8690	0.8650	0.9050	0.9150
2	1	0.5	0.8160	0.8720	0.8800	0.8760	0.9220	0.9270
2	1	0.9	0.8660	0.8910	0.8800	0.9200	0.9400	0.9330
2	2	0.1	0.8010	0.8700	0.8720	0.8660	0.9110	0.9250
2	2	0.5	0.8580	0.8990	0.8830	0.9210	0.9440	0.9350
2	2	0.9	0.8660	0.8810	0.8720	0.9290	0.9310	0.9330
8	1	0.1	0.7340	0.8090	0.8620	0.8100	0.8770	0.9190
8	1	0.5	0.8210	0.8620	0.9130	0.8760	0.9150	0.9540
8	1	0.9	0.8650	0.8790	0.8850	0.9160	0.9300	0.9310
8	2	0.1	0.7400	0.8320	0.8640	0.8130	0.8950	0.9080
8	2	0.5	0.7760	0.9610	0.9030	0.8390	0.9800	0.9460
8	2	0.9	0.7480	0.9960	0.8790	0.8300	0.9980	0.9330
32	1	0.1	0.7470	0.8180	0.8750	0.8100	0.8820	0.9200
32	1	0.5	0.8330	0.8660	0.9180	0.8930	0.9280	0.9650
32	1	0.9	0.8460	0.8710	0.8790	0.9190	0.9330	0.9430
32	2	0.1	0.7250	0.8350	0.8800	0.7950	0.9010	0.9310
32	2	0.5	0.7130	0.9460	0.8980	0.8040	0.9670	0.9430
32	2	0.9	0.7030	0.9980	0.8690	0.7770	0.9990	0.9420

Table 5: Coefficient estimates and standard errors, Model Setup A

	Full	Equal	AIC	BIC	S-AIC	S-BIC	JMA	Plug-In	WALS
CONSTANT	0.0609	0.0603	0.0518	0.0441	0.0526	0.0474	0.0559	0.0641	0.0594
	(0.0193)	(0.0192)	(0.0214)	(0.0210)	(0.0200)	(0.0204)	(0.0201)	(0.0182)	(0.0221)
GDP60	-0.0155	-0.0157	-0.0145	-0.0138	-0.0144	-0.0135	-0.0156	-0.0156	-0.0156
	(0.0030)	(0.0028)	(0.0031)	(0.0031)	(0.0030)	(0.0030)	(0.0029)	(0.0027)	(0.0033)
EQUIPINV	0.1366	0.1835	0.1377	0.1518	0.1501	0.1686	0.1511	0.2263	0.1555
	(0.0400)	(0.0361)	(0.0397)	(0.0394)	(0.0383)	(0.0363)	(0.0390)	(0.0349)	(0.0551)
SCHOOL60	0.0170	0.0173	0.0191	0.0157	0.0168	0.0157	0.0181	0.0137	0.0175
	(0.0085)	(0.0081)	(0.0081)	(0.0082)	(0.0082)	(0.0081)	(0.0081)	(0.0085)	(0.0097)
LIFE60	0.0008	0.0009	0.0008	0.0009	0.0008	0.0008	0.0009	0.0010	0.0009
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)
DPOP	0.3466	0.1736	0.3275	0.1240	0.2433	0.1367	0.2465	0.0055	0.2651
	(0.1911)	(0.1706)	(0.1853)	(0.1797)	(0.1784)	(0.1699)	(0.1760)	(0.1718)	(0.2487)
LAW	0.0174	0.0094	0.0167	0.0154	0.0142	0.0097	0.0166		0.0147
	(0.0058)	(0.0028)	(0.0056)	(0.0057)	(0.0049)	(0.0034)	(0.0052)		(0.0065)
TROPICS	-0.0075	-0.0040	-0.0083		-0.0052	-0.0029	-0.0043		-0.0055
	(0.0036)	(0.0018)	(0.0036)		(0.0023)	(0.0013)	(0.0018)		(0.0037)
AVELF	-0.0077	-0.0048			-0.0033	-0.0015	-0.0026	-0.0104	-0.0053
	(0.0066)	(0.0033)			(0.0026)	(0.0011)	(0.0016)	(0.0065)	(0.0048)
CONFUC	0.0562	0.0317	0.0596	0.0627	0.0600	0.0633	0.0430	0.0251	0.0443
	(0.0129)	(0.0062)	(0.0129)	(0.0129)	(0.0126)	(0.0123)	(0.0088)	(0.0045)	(0.0163)

Note: Standard errors are reported in parentheses. The column labeled WALS displays the weighted-average least-squares estimates of Magnus, Powell, and Prufer (2010, Table 2).

Table 6: Coefficient estimates and standard errors, Model Setup B

	Full	Equal	AIC	BIC	S-AIC	S-BIC	JMA	Plug-In	WALS
CONSTANT	0.0609	0.0254	0.0674	0.0344	0.0556	0.0452	0.0526	0.0734	0.0560
	(0.0193)	(0.0097)	(0.0182)	(0.0138)	(0.0156)	(0.0140)	(0.0146)	(0.0106)	(0.0215)
GDP60	-0.0155	-0.0060	-0.0146	-0.0120	-0.0138	-0.0126	-0.0137	-0.0153	-0.0136
	(0.0030)	(0.0011)	(0.0031)	(0.0029)	(0.0028)	(0.0027)	(0.0025)	(0.0018)	(0.0033)
EQUIPINV	0.1366	0.1094	0.1484	0.1951	0.1510	0.1593	0.1322		0.1037
	(0.0400)	(0.0170)	(0.0390)	(0.0340)	(90.0338)	(0.0300)	(0.0206)		(0.0537)
SCHOOL60	0.0170	0.0115	0.0203		0.0117	0.0066	0.0139		0.0125
	(0.0085)	(0.0033)	(0.0080)		(0.0043)	(0.0021)	(0.0026)		(0.0094)
LIFE60	0.0008	0.0004	0.0006	0.0012	0.0008	0.0010	0.0008	0.0010	0.0008
	(0.0003)	(0.0001)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0003)
DPOP	0.3466	0.0607			0.0666	0.0136	0.1804		0.2236
	(0.1911)	(0.0717)			(0.0471)	(0.0141)	(0.0707)		(0.2156)
LAW	0.0174	0.0092	0.0140		0.0119	0.0076	0.0151	0.0171	0.0137
	(0.0058)	(0.0022)	(0.0053)		(0.0040)	(0.0024)	(0.0032)	(0.0031)	(0.0063)
TROPICS	-0.0075	-0.0037	-0.0064		-0.0034	-0.0015	-0.0042	-0.0032	-0.0055
	(0.0036)	(0.0017)	(0.0034)		(0.0017)	(0.0008)	(0.0017)	(0.0025)	(0.0039)
AVELF	-0.0077	-0.0040			-0.0036	-0.0018	-0.0034	-0.0091	-0.0083
	(0.0066)	(0.0032)			(0.0026)	(0.0012)	(0.0017)	(0.0044)	(0.0057)
CONFUC	0.0562	0.0419	0.0616	0.0728	0.0640	0.0688	0.0444		0.0451
	(0.0129)	(0.0057)	(0.0128)	(0.0120)	(0.0123)	(0.0121)	(0.0080)		(0.0163)

Note: Standard errors are reported in parentheses. The column labeled WALS displays the weighted-average least-squares estimates of Magnus, Powell, and Prufer (2010, Table 3).

Table 7: Weights placed on each submodel, Model Setup A

Model	AIC	BIC	JMA	Plug-In
4	0.000	0.000	0.070	0.000
5	0.000	0.000	0.000	0.624
8	0.000	0.000	0.243	0.000
9	0.000	0.000	0.071	0.000
10	0.000	1.000	0.424	0.000
12	1.000	0.000	0.192	0.000
13	0.000	0.000	0.000	0.376

Table 8: Weights placed on each submodel, Model Setup B

Model	AIC	BIC	JMA	Plug-In
1	0.000	0.000	0.000	0.300
72	0.000	0.000	0.087	0.000
168	0.000	0.000	0.269	0.000
234	0.000	0.000	0.000	0.700
259	0.000	0.000	0.026	0.000
268	0.000	1.000	0.190	0.000
296	0.000	0.000	0.033	0.000
368	1.000	0.000	0.000	0.000
378	0.000	0.000	0.394	0.000

Table 9: Regressor set of the submodel, Model Setup A

Model	Regressor Set
4	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+LAW+TROPICS
5	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+AVELF
8	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+LAW+TROPICS+AVELF
9	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+CONFUC
10	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+LAW+CONFUC
12	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+LAW+TROPICS+CONFUC
13	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+DPOP+AVELF+CONFUC

Table 10: Regressor set of the submodel, Model Setup B

Model	Regressor Set
1	CONSTANT
72	CONSTANT+GDP60+EQUIPINV+SCHOOL60+TROPICS
168	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LAW+AVELF
234	CONSTANT+GDP60+LIFE60+LAW+TROPICS+AVELF
259	CONSTANT+EQUIPINV+CONFUC
268	CONSTANT+GDP60+EQUIPINV+LIFE60+CONFUC
296	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LAW+CONFUC
368	CONSTANT+GDP60+EQUIPINV+SCHOOL60+LIFE60+LAW+TROPICS+CONFUC
378	CONSTANT+GDP60+LIFE60+DPOP+LAW+TROPICS+CONFUC

References

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, ed. by B. Petroc, and F. Csake, pp. 267–281. Akademiai Kiado.

Andrews, D. (1991a): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

——— (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Buckland, S., K. Burnham, and N. Augustin (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

Burnham, K., and D. Anderson (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Verlag.

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

DiTraglia, F. (2011): "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," *Unpublished Manuscript*.

Durlauf, S., A. Kourtellos, and C. Tan (2008): "Are Any Growth Theories Robust?," *The Economic Journal*, 118(527), 329–346.

Durlauf, S. N., P. A. Johnson, and J. R. Temple (2005): "Growth Econometrics," in *Handbook of Economic Growth*, ed. by P. Aghion, and S. Durlauf, vol. 1, pp. 555–677. Elsevier.

Fernandez, C., E. Ley, and M. Steel (2001): "Model Uncertainty in Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16(5), 563–576.

Hansen, B. (2007): "Least Squares Model Averaging," *Econometrica*, 75(4), 1175–1189.

——— (2009): "Averaging Estimators for Regressions with a Possible Structural Break," *Econometric Theory*, 25(06), 1498–1514.

——— (2010): "Averaging Estimators for Autoregressions with a Near Unit Root," *Journal of Econometrics*, 158(1), 142–155.

Hansen, B., and J. Racine (2012): "Jackknife Model Averaging," *Journal of Econometrics*, 167(1), 38–46.

Hansen, P., A. Lunde, and J. Nason (2011): "The Model Confidence Set," *Econometrica*, 79(2), 453–497.

Hausman, J. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

Hjort, N., and G. Claeskens (2003): "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98(464), 879–899.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401.

Kabaila, P. (1995): "The Effect of Model Selection on Confidence Regions and Prediction Regions," *Econometric Theory*, 11, 537–537.

——— (1998): "Valid Confidence Intervals in Regression after Variable Selection," *Econometric Theory*, 14(4), 463–482.

Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *The Annals of Statistics*, 18, 191–219.

Leeb, H., and B. Pötscher (2003): "The Finite-Sample Distribution of Post-Model-Selection Estimators and Uniform versus Non-Uniform Approximations," *Econometric Theory*, 19(1), 100–142.

——— (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21(1), 21–59.

——— (2006): "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?," *The Annals of Statistics*, 34(5), 2554–2591.

——— (2008): "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?," *Econometric Theory*, 24(02), 338–376.

Leung, G., and A. Barron (2006): "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52(8), 3396–3410.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Liang, H., G. Zou, A. Wan, and X. Zhang (2011): "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106, 1053–1066.

Magnus, J., O. Powell, and P. Prufer (2010): "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics," *Journal of Econometrics*, 154(2), 139–153.

Newey, W., and K. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Pötscher, B. (1991): "Effects of Model Selection on Inference," *Econometric Theory*, 7(2), 163–185.

——— (2006): "The Distribution of Model Averaging Estimators and an Impossibility Result Regarding its Estimation," *Lecture Notes-Monograph Series*, 52, 113–129.

Sala-i Martin, X., G. Doppelhofer, and R. Miller (2004): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 813–835.

Staiger, D., and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer Verlag.

Wan, A., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156(2), 277–283.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

——— (1984): *Asymptotic Theory for Econometricians*. Academic Press.

White, H., and X. Lu (2010): "Robustness Checks and Robustness Tests in Applied Economics," *Unpublished Manuscript*.

Yang, Y. (2001): "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96(454), 574–588.

Yuan, Z., and Y. Yang (2005): "Combining Linear Regression Models: When and How?," *Journal of the American Statistical Association*, 100, 1202–1214.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476), 1418–1429.