Communications for Statistical Applications and Methods
2021, Vol. 28, No. 2, 189–204

https://doi.org/10.29220/CSAM.2021.28.2.189
Print ISSN 2287-7843 / Online ISSN 2383-4757

Robustness of model averaging methods for the violation of standard linear regression assumptions

Yongsu Lee[1,a], Juwon Song[b]

[a]Department of Statistics, University of Wisconsin-Madison, USA;
[b]Department of Statistics, Korea University, Korea

Abstract

In a regression analysis, a single best model is usually selected among several candidate models. However, it is often useful to combine several candidate models to achieve better performance, especially, in the prediction viewpoint. Model combining methods such as stacking and Bayesian model averaging (BMA) have been suggested from the perspective of averaging candidate models. When the candidate models include a true model, it is expected that BMA generally gives better performance than stacking. On the other hand, when candidate models do not include the true model, it is known that stacking outperforms BMA. Since stacking and BMA approaches have different properties, it is difficult to determine which method is more appropriate under other situations. In particular, it is not easy to find research papers that compare stacking and BMA when regression model assumptions are violated. Therefore, in the paper, we compare the performance among model averaging methods as well as a single best model in the linear regression analysis when standard linear regression assumptions are violated. Simulations were conducted to compare model averaging methods with the linear regression when data include outliers and data do not include them. We also compared them when data include errors from a non-normal distribution. The model averaging methods were applied to the water pollution data, which have a strong multicollinearity among variables. Simulation studies showed that the stacking method tends to give better performance than BMA or standard linear regression analysis (including the stepwise selection method) in the sense of risks (see (3.1)) or prediction error (see (3.2)) when typical linear regression assumptions are violated.

Keywords: model averaging, stacking regression, Bayesian model averaging, outliers, misspecified distribution

1. Introduction

When we conduct a regression analysis, it is common to choose one model which best explains the response variable. On the other hand, there exist the cases when the response variable can be better predicted by combining several plausible models. Both *Staking* and *Bayesian model averaging* (hereafter, BMA) (Madigan and Raftery, 1994; Wolpert, 1992) use a concept of combining candidate models with a carefully chosen *candidate model weights*. In the stacking, weights are chosen to minimize the sum of the squares of the distances between response variable values and a linear combination of predicted values obtained by each candidate model. On the other hand, BMA takes advantage of the posterior probability to combine candidate models, and the weights are selected by the relative contribution of each model compared to other candidate models using approximated Bayesian information criterion (BIC).

[1] Corresponding author: Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA.
E-mail: yongsulee@stat.wisc.edu

Published 31 March 2021 / journal homepage: http://csam.or.kr

Even though the stacking and BMA share the similarity in the viewpoint of fusing candidate models to get a better model using model combining weights, each method is based on different types of combining weights and, consequently, they may show dissimilar results in prediction or accuracy of parameter-estimates under various circumstances. When candidate models include a true model, BMA gives a better performance than stacking or, at least, provides performance as much as stacking in perspective of the distance between the regression coefficient parameters and its estimates ($\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|^2$, the risk) (Clarke, 2003). In contrast, when candidate models do not include a true model, it is known that the stacking outperforms BMA in prediction or estimating coefficients (Clarke, 2003).

The idea of combining several candidate models can be considered as "ensemble learnings" in general, and many ensemble learning methods have been suggested under various frameworks. In order to improve performance in classification problems, *bagging* (Breiman, 1996a) and *random forest* method(Breiman, 2001) were introduced, especially for tree models, and Schapire (2003) proposed *boosting*, which combines trees based on weighted samples. For the regression problems, stacking and BMA are two representative methods. The stacking was originally proposed by Wolpert (1992) and has been developed by Breiman (1996b). Hjort and Claeskens (2003) suggested a more sophisticated version of the stacking method, say "Frequentist Model Averaging (FMA)", by establishing a large-sample likelihood framework to accurately describe properties of model averaging estimators. An optimal weight choice for the FMA has been studied by Liang *et al.* (2011). For a sparse model averaging, Ando and Li (2014) proposed a methodology of selecting model averaging weights under a high-dimensional regression setting, and it was also extended to the generalized linear models case (Ando *et al.*, 2017). Recently, Zhang *et al.* (2020) studied asymptotic distributions of weights for the model averaging when the number of parameter diverges. On the other hand, BMA approach has been developed by Madigan and Raftery (1994) and Kass and Raftery (1995), and Hoeting *et al.* (1999) provided elaborate reviews of BMA applications. A study of priors for BMA can be found in Fernandez *et al.* (2001), and Eklund and Karlsson (2007) suggested a new approach of BMA using predictive likelihoods. Note that most of studies have been focusing on either suggesting a new framework of selecting weights under different circumstances or studying an (asymptotic) theoretical properties of the estimator of weights. However, since there exist various versions of the model averaging methods, it would be practically useful if we have a guideline of choosing the best model averaging method according to a specific situation.

In the standard regression analysis, in which necessary assumptions should be satisfied, it is often the case that we have a dataset which does not satisfy one (or many) of the assumptions. Detecting such a violation sometimes is not straightforward even with rigorous regression diagnostics. Furthermore, it could be challenging to take an action to rectify such a violation if we do not have enough information about the given dataset or the true model. Therefore, it is worthwhile to investigate a way of providing a robust statistical model even though the standard linear regression assumptions are violated. The purpose of the study is to compare the results of model averaging techniques with traditional regression model when several model assumptions are not met. We conduct several simulation studies under different scenarios. Each scenario is designed to violate one of standard linear regression assumptions, and we compare the model averaging results with one from traditional linear regression model including the model selected by the stepwise method.

The remainder of the paper is organized as follows. In Section 2, we introduce the stacking method and Bayesian model averaging, and discuss how to construct a list of candidate models for the model averaging methods. In Section 3, several simulation studies are conducted to compare performance of the stacking method and BMA as well as the result of the linear regression with full model and a model constructed by the stepwise selection. The simulation deals with two scenarios: (1)

fitting a regression model with a dataset containing many outliers, and (2) fitting a regression model when a distributional assumption is violated. Here, we consider a dataset whose error terms come from a heavy-tailed distribution or whose response variable actually follows a Poisson distribution. In Section 4, we compare performance among aforementioned methods using the water pollution dataset. Finally, Section 5 summarizes the findings and discusses a direction of further studies.

2. Background: Model averaging methods

2.1. Problem formulation

Suppose that we have p predictor variables $\boldsymbol{X} = (X_1, \ldots, X_p)$ to explain a response variable Y, and define $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \in \mathbb{R}^{n\times 1}$ as corresponding collections of n data points, respectively. Denote $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^{p\times 1}$ be the i^{th} observation of $\boldsymbol{X}$ (or i^{th} row of $\mathbf{X}$), and, similarly, let $y_i \in \mathbb{R}$ be the i^{th} datapoint of Y.

A standard linear regression model is expressed by,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \text{where } \varepsilon \sim N\left(0, \sigma^2\right), \tag{2.1}$$

and let us consider this as a true model. Let $\hat{f}(\mathbf{x}_i)$ be an estimator of the true model, $f(\boldsymbol{X}) = E[Y|\boldsymbol{X}]$. For the i^{th} observation $\mathbf{x}_i$, that is,

$$\hat{f}(\mathbf{x}_i) := \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \tag{2.2}$$

where $\hat{\beta}_k,\ k = 0, 1, \ldots, p$, is an estimator of β_k based on the datapoints $\mathbf{X}$ and $\mathbf{y}$.

The *model averaging* combines several models instead of selecting a single model, and, consequently, we need to take account of a set of properly selected *candidate models*. Suppose that we have a list of candidate models, $\boldsymbol{\mathcal{M}}$, consisting of M candidate models, $\boldsymbol{\mathcal{M}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. In the standard linear regression model, various combinations of predictor variables can be considered as a list of candidate models, that is, $\mathcal{M}_m \subseteq \{X_1, X_2, \ldots, X_p\}$, for $m = 1, \ldots, M$. For example, $\mathcal{M}_1 = \{X_1, X_2, X_3\}$ represents a candidate (linear regression) model using X_1, X_2, X_3 only. Given a candidate model, $\mathcal{M}_m$, define an estimator of $f(\mathbf{x})$ as

$$\hat{f}_m(\mathbf{x}_i) = \hat{\gamma}_0^{(m)} + \hat{\gamma}_1^{(m)} x_{i1} + \cdots + \hat{\gamma}_p^{(m)} x_{ip}, \tag{2.3}$$

where for $m = 1, \ldots M$ and $k = 1, \ldots, p$,

$$\hat{\gamma}_k^{(m)} = \begin{cases} \hat{\beta}_k^{(m)}, & \text{if } k \in \mathcal{M}_m, \\ 0, & \text{otherwise,} \end{cases} \qquad \hat{\gamma}_0^{(m)} = \begin{cases} \hat{\beta}_0^{(m)}, & \text{if } \mathcal{M}_m \text{ includes an intercept,} \\ 0, & \text{otherwise.} \end{cases}$$

$\hat{\beta}_k^{(m)},\ k = 1, \ldots, p$ is the ordinary least square (OLS) estimator of the regression coefficient of X_k in $\mathcal{M}_m$, and $\hat{\beta}_0^{(m)}$ indicates the OLS estimator of the intercept in $\mathcal{M}_m$. Therefore, we can construct a matrix containing all the OLS estimators across M candidate models such that,

$$\hat{\boldsymbol{\Gamma}} = \begin{bmatrix} \hat{\gamma}_1^{(1)} & \cdots & \hat{\gamma}_1^{(M)} \\ \vdots & \ddots & \vdots \\ \hat{\gamma}_p^{(1)} & \cdots & \hat{\gamma}_p^{(M)} \\ \hat{\gamma}_0^{(1)} & \cdots & \hat{\gamma}_0^{(M)} \end{bmatrix}, \qquad \hat{\boldsymbol{\Gamma}} \in \mathbb{R}^{(p+1)\times M}. \tag{2.4}$$

Note that some of elements of m^{th} column of $\hat{\boldsymbol{\Gamma}}$ are zeros if corresponding predictors are not included for the candidate model $\mathcal{M}_m$.

2.2. Stacking

Stacking method (or *Stacked Generalization*, Wolpert (1992)) is a model averaging technique from a frequentist viewpoint (Friedman *et al.*, 2001). When it comes to a standard linear regression model, the stacking provides an improved prediction accuracy by forming a linear combination of different candidate models (Breiman, 1996b). Specifically, in the stacking method, coefficients are chosen to minimize the sum of squared of differences between values of the response variable and a linear combination of predicted values from fitted regression model based on different candidate models.

Since the combining models based on the whole training dataset may cause a problem of overfitting in the stacking, the *leave-one-out* method has been considered when obtaining predicted values. Let $\hat{f}_m^{-i}(\mathbf{x})$ be a fitted model using predictors in $\mathcal{M}_m$ based on the training data in which i^{th} observation is eliminated. Thus, a possible way to achieve a stacking weight can be formulated by an optimization program

$$\hat{\boldsymbol{\alpha}} = \text{argmin}_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left[y_i - \sum_{m=1}^{M} \alpha_m \hat{f}_m^{-i}(\mathbf{x}_i) \right]^2, \tag{2.5}$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_M)^\top$ (Friedman *et al.*, 2001). Using $\hat{\boldsymbol{\alpha}}$, we can construct a stacking estimator

$$\hat{f}^{\text{st}}(\mathbf{x}_i) := \sum_{m=1}^{M} \hat{\alpha}_m \hat{f}_m(\mathbf{x}_i) \tag{2.6}$$

based on a given M candidate models. That is, stacking estimators of coefficients and intercept term in (2.2) can be expressed as

$$\begin{aligned} \hat{\beta}_0^{st} &:= \hat{\alpha}_1 \hat{\gamma}_0^{(1)} + \cdots + \hat{\alpha}_M \hat{\gamma}_0^{(M)} \\ \hat{\beta}_k^{st} &:= \hat{\alpha}_1 \hat{\gamma}_k^{(1)} + \cdots + \hat{\alpha}_M \hat{\gamma}_k^{(M)}, \qquad k = 1, \ldots, p, \end{aligned}$$

or simply, by using (2.4) and (2.5), where $\hat{\boldsymbol{\beta}}^{st} = (\hat{\beta}_1^{st}, \ldots, \hat{\beta}_p^{st}, \hat{\beta}_0^{st})^\top$,

$$\hat{\boldsymbol{\beta}}^{st} = \hat{\boldsymbol{\Gamma}}\, \hat{\boldsymbol{\alpha}}.$$

Various determinations of the combining coefficients can be obtained by putting a constraint on the weights. The five types of restriction are often considered to compare a performance of the stacking method. The first model (S1) does not impose any constraint on the weight, and Clarke (2003) suggests three types of restrictions to improve performance of the stacking method:

- Restriction Type 1 (S2): $\sum_{m=1}^{M} \alpha_m = 1$,
- Restriction Type 2 (S3): From S2, by replacing negative α_m with zero and then re-normalizing so the nonzero α_m's sum to one.
- Restriction Type 3 (S4): $\sum_m \alpha_m = 1$, $\alpha_m \geq 0$ for all $m = 1, \ldots, M$.

Also, we consider an additional restriction type, which gives competitive performance empirically:

- Restriction Type 4 (S5): $\alpha_m \geq 0$ for all $m = 1, \ldots, M$.

Note that S2 (or S4) is *not* obtained by normalizing S1 (or S5), because each restriction is considered at the optimization process.

2.3. Bayesian model averaging

To construct a parsimonious model, we usually select significant (explanatory) variables based on the p-values. However, such a process may mislead the result by selecting inappropriate variables against the intuition or disregarding model uncertainty (Raftery, 1995). BIC has been suggested as an alternative way for model selection. BMA method based on the BIC suggests a reasonable framework to compensate the model uncertainty.

In the linear regression model (2.1), we can express a posterior probability of $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p)^\top$ given the data $\mathbf{x}$ as

$$P[\boldsymbol{\beta}|\mathbf{x}] = \sum_{m=1}^{M} P(\boldsymbol{\beta}|\mathcal{M}_m, \mathbf{x}) \cdot P(\mathcal{M}_m|\mathbf{x}). \tag{2.7}$$

In the linear regression model, the BIC approximation plays an important role to approximate $P(\mathcal{M}_m|\mathbf{x})$ in (2.7) (Raftery, 1995) as follows:

$$P[\mathcal{M}_m|\mathbf{x}] \approx \frac{\exp\left\{-\frac{1}{2} \cdot \mathrm{BIC}'_m\right\}}{\sum_{\ell=1}^{M} \exp\left\{-\frac{1}{2} \cdot \mathrm{BIC}'_\ell\right\}}, \quad m = 1, \ldots, M, \tag{2.8}$$

$$\mathrm{BIC}'_m = N \cdot \log\left(1 - R_m^2\right) + p_m \cdot \log N, \tag{2.9}$$

where R_m^2 is the coefficient of determination for the fitted linear regression model using the predictors in $\mathcal{M}_m$, and p_m indicates the number of predictors in the model $\mathcal{M}_m$.

Under the given candidate models, $\mathcal{M}_1, \ldots, \mathcal{M}_M$, estimators of regression coefficients can be derived by using their posterior means such that, for $k = 1, \ldots, p$,

$$\begin{aligned} \hat{\beta}_k^{\mathrm{BMA}} &= E[\beta_k|\mathbf{x}] \\ &= \sum_{m=1}^{M} E[\beta_k|\mathcal{M}_m, \mathbf{x}] \cdot P[\mathcal{M}_m|\mathbf{x}]. \end{aligned} \tag{2.10}$$

Practically, $E[\beta_k|\mathcal{M}_m, \mathbf{x}]$ in (2.10) under the candidate model $\mathcal{M}_m$ can be estimated by its maximum likelihood estimator (MLE), $\tilde{\beta}_k^{(m)}$, and combining it with the approximation from (2.8) gives

$$\hat{\beta}_k^{\mathrm{BMA}} \approx \sum_{m=1}^{M} \tilde{\beta}_k^{(m)} \cdot \frac{\exp\left\{-\frac{1}{2} \cdot \mathrm{BIC}'_m\right\}}{\sum_{\ell=1}^{M} \exp\left\{-\frac{1}{2} \cdot \mathrm{BIC}'_\ell\right\}}. \tag{2.11}$$

2.4. Selecting candidate models

Selecting a list of candidate models is one of the essential part of the model averaging method because the performance of the method heavily depends on the selected candidate models. For example, if a list consists of implausible models to explain the data, the performance of the averaging method would be even worse than the ordinary regression analysis. However, there exist few of studies to choose ideal candidate models with a solid theoretical support.

Breiman (1996b) suggested choosing p candidate models for p dimensional data. Specifically, suppose that we have a list of candidate models $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. A candidate model $\mathcal{M}_m$ has m predictors for $m = 1, \ldots, M\,(= p)$. The m predictors can be selected by comparing the R^2 (coefficient of determination) among all the possible linear regression model with m predictors. However, when p

gets larger, there exist 2^p possible candidate models to examine, which possibly results in an inefficient process to select candidate models. The *leaps-and-bound* algorithm (Furnival and Wilson, 1974) can be used to choose the best subset regression models more efficiently.

In Bayesian approach, a model selection method based on BIC, called *Occam's Window* (Raftery, 1995), has been proposed. However, when p is still too large, it is not efficient to use the Occam's Window method. For this case, preprocessing with the leaps-and-bound algorithm can be a useful way of selecting candidate models before applying the Occam's Window to construct a list of candidate models.

3. Simulation studies

Two possible scenarios violating standard linear regression assumptions are considered to verify performance of the model averaging methods and compare them with the result based on the standard linear regression analysis. The performance of each method has been evaluated by two criteria: (1) Risk (Clarke, 2003) is defined by

$$\text{Risk} := \left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right\|; \tag{3.1}$$

(2) Prediction Error (Breiman, 1996b) is calculated by

$$\text{PE} := \| \mathbf{y}_{\text{test}} - \hat{\mathbf{y}}_{\text{test}} \|. \tag{3.2}$$

Note that we generate a test dataset, $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$, independently, and $(\hat{\mathbf{y}}_{\text{test}})_i$ is a fitted value, $\hat{f}(\mathbf{x}_{\text{test},i})$ where $\mathbf{x}_{\text{test},i}$ is the i^{th} row (observation) of $\mathbf{X}_{\text{test}}$. The evaluation criteria were calculated for eight different methods for all the simulations. For the stacking, we fit five methods with different types of restrictions (Section 2.2), and denote them as S1, S2, S3, S4, and S5 respectively. Similarly, BMA is considered as the another model averaging technique (Section 2.3). As the standard linear regression analysis, we consider the full model (Full) and the model chosen from the stepwise selection method (Step).

3.1. When a dataset contains outliers

If data contains some outliers, it may violate the standard assumptions in a linear regression model, and, consequently, a result of analysis may lead to an unexpected conclusion for the prediction or estimates of regression coefficients. The following simulation is designed to confirm whether the model averaging method is useful to reduce the PE or Risk in this situation.

We set a true regression model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{20} X_{20} + \varepsilon. \tag{3.3}$$

Two hundred observations of $X_1, X_2, \ldots, X_{20}$ were generated from a standard normal distribution, independently. The regression coefficients, β_j, $j = 1, 2, \ldots, 20$, and an intercept term β_0 are set as:

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
$C1$:	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
$C2$:	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0
$C3$:	1	0	0	0	0	16	0	0	0	0	16	0	0	0	0	16	0	0	0	0	0
$C4$:	1	0	0	1	4	9	16	8	4	1	1	4	9	16	9	4	1	1	0	0	0

(3.4)

Table 1: Averages of evaluation criteria over the 100 simulations when a dataset generated with errors from the standard normal distribution (no outlier). The true model is given in (3.3), C1, . . . , C4 represent different sets of coefficients (See (3.4)), and 100 observations are considered for each simulation.

Risks (No outlier)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
C1	1.9402	0.9983	2.0682	2.0625	1.6434	0.8975	0.8955	**0.6042**
(s.e.)	(0.3704)	(0.4077)	(0.3805)	(0.3794)	(0.3780)	(0.3118)	(0.3112)	(0.2428)
C2	1.8690	1.6590	2.2220	2.2127	3.1281	1.6899	1.6882	**1.5923**
(s.e.)	(0.3558)	(0.3795)	(0.4227)	(0.4228)	(1.4831)	(0.3367)	(0.3333)	(0.3343)
C3	1.8836	0.9886	2.0187	2.0058	1.6240	0.8807	0.8802	**0.5912**
(s.e.)	(0.3304)	(0.4169)	(0.3588)	(0.3655)	(0.3629)	(0.2967)	(0.2964)	(0.2406)
C4	1.8098	1.5323	2.0761	2.0769	1.9518	1.5201	1.5191	**1.4685**
(s.e.)	(0.2788)	(0.2816)	(0.3627)	(0.3581)	(0.5592)	(0.2679)	(0.2701)	(0.2562)
Prediction errors (No outlier)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
C1	89.65	85.90	90.98	90.83	88.18	83.74	83.71	**82.21**
(s.e.)	(7.58)	(8.34)	(8.41)	(8.41)	(7.84)	(7.57)	(7.57)	(7.29)
C2	90.47	89.65	93.98	93.87	108.97	89.75	89.72	**88.92**
(s.e.)	(7.49)	(7.23)	(8.12)	(8.22)	(24.21)	(7.31)	(7.29)	(7.15)
C3	90.29	86.87	91.16	91.01	88.97	84.80	84.79	**83.45**
(s.e.)	(7.85)	(7.84)	(7.93)	(8.03)	(7.78)	(6.98)	(6.98)	(6.57)
C4	88.25	87.51	90.85	90.83	90.67	87.10	87.10	**86.66**
(s.e.)	(7.13)	(7.29)	(7.33)	(7.25)	(9.30)	(7.00)	(6.99)	(7.01)

A set of coefficients in C1 corresponds to a model that contains three nonzero (significant) coefficients. C3 is similarly constructed as C1, while C3 is chosen to check the effect of a few large coefficients in the model averaging methods. We also consider the coefficients which have fewer number of zero coefficients. For example, C2 contains twenty coefficients that consist of sixteen ones and four zeros. C4 indicates a similar model with C2, but it includes more various magnitudes of coefficients. 100 datapoints $(x_{i1}, \ldots, x_{i,100})$ are generated from the standard normal distribution, independently. We also generate 100 error terms ε_i from the normal distribution, but, in order to consider data with outliers, we add the five fold of the range of $\{\varepsilon_i\}$ (that is, $\delta := 5 \cdot (\max_i \varepsilon_i - \min_i \varepsilon_i)$) to the randomly selected 20 error terms (20% of observations) considering their signs. In other words, if ε_i, $i = 1, \ldots, 20$, is positive, we add δ to the error term, and otherwise we subtract δ from the error term. For the comparison purpose, we generate two types of datasets: (1) a dataset with error terms from the standard normal distribution; (2) a dataset with modified errors term designed as outliers as previously described. For both types of datasets, y_i is calculated from (3.3), and a list of candidate models are selected based on the method suggested by Breiman (1996b).

Table 1 and Table 2 show the average of evaluation metrics using datasets with no outliers and with some outliers, respectively, over hundred repetitions of the simulation. The boldface numbers in Table 1 and Table 2 indicate the best method with the minimum average risks and prediction errors for each set of coefficients. Boxplots for each simulation result are also provided in Figure 1 and Figure 2. When a dataset does not include outliers (Table 1), almost all the averages of risks for model averaging methods (S4, S5, BMA) are smaller than the full model and the model based on selected variables using stepwise selection. Particularly, BMA shows consistently better performance for all the coefficients cases than any other method. When the data contain 20% outliers, BMA shows better performance in C1 and C3. In contrast, in C2 and C4 cases (and even C1 for PEs), the stacking method (S5, and some of S4 cases) outperforms not only Full and Step, but also BMA. This can also

Table 2: Averages of evaluation criteria over the 100 simulations when a dataset designed to contain outliers (20% of observations). The true model is given in (3.3), C1, . . . , C4 represent different sets of coefficients (See (3.4)), and 100 observations are considered for each simulation.

Risks (with many Outliers)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
C1	21.6586	10.8459	22.8962	23.3523	17.9020	10.0628	9.0657	**7.2889**
(s.e.)	(4.5460)	(4.4022)	(5.0095)	(5.1169)	(4.3799)	(3.1255)	(3.1347)	(2.0522)
C2	21.9146	21.4316	25.6144	25.9401	20.3284	17.6092	**16.7498**	17.8946
(s.e.)	(4.2928)	(3.1270)	(5.6014)	(5.6299)	(3.8008)	(2.6933)	(2.7304)	(1.8370)
C3	20.5704	9.5758	22.7116	22.7149	17.3891	8.8578	8.8182	**6.0188**
(s.e.)	(3.9570)	(4.0545)	(4.9428)	(4.8278)	(3.6904)	(3.1269)	(3.1269)	(2.6406)
C4	21.8171	19.8696	26.5458	26.3956	20.6527	18.7189	**18.7085**	18.7497
(s.e.)	(4.3304)	(4.3672)	(5.3940)	(5.4194)	(4.3079)	(3.9336)	(3.9245)	(4.0056)
Prediction errors (with many Outliers)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
C1	902.23	814.33	916.09	925.53	865.52	777.34	**745.55**	747.25
(s.e.)	(107.26)	(110.46)	(110.26)	(113.47)	(102.76)	(97.31)	(94.43)	(94.84)
C2	899.82	875.05	963.67	969.65	870.73	823.95	**803.73**	822.21
(s.e.)	(100.65)	(91.67)	(117.19)	(119.90)	(95.86)	(86.85)	(83.91)	(80.74)
C3	873.92	789.53	906.88	907.11	845.81	745.83	744.99	**709.13**
(s.e.)	(95.08)	(104.53)	(108.29)	(107.76)	(89.81)	(88.32)	(88.24)	(88.71)
C4	898.91	888.89	976.64	975.89	884.48	870.04	**869.72**	882.37
(s.e.)	(105.06)	(107.31)	(100.97)	(100.77)	(100.83)	(100.46)	(100.68)	(96.75)

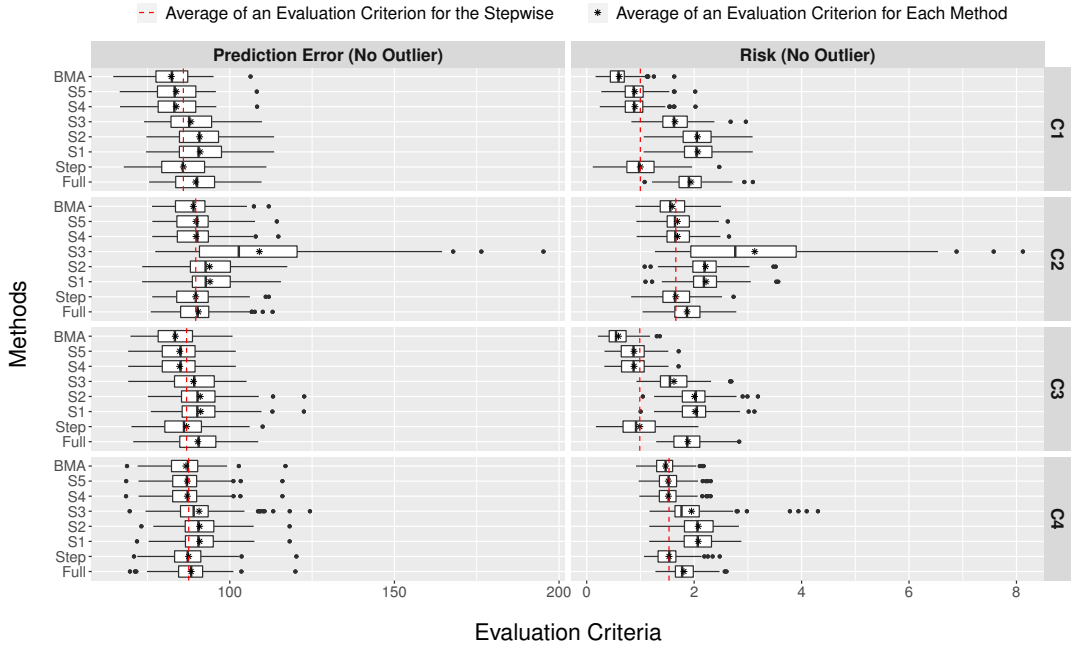

Figure 1: *Boxplots of PEs and Risks over the 100 simulations when a dataset does not include outliers.*

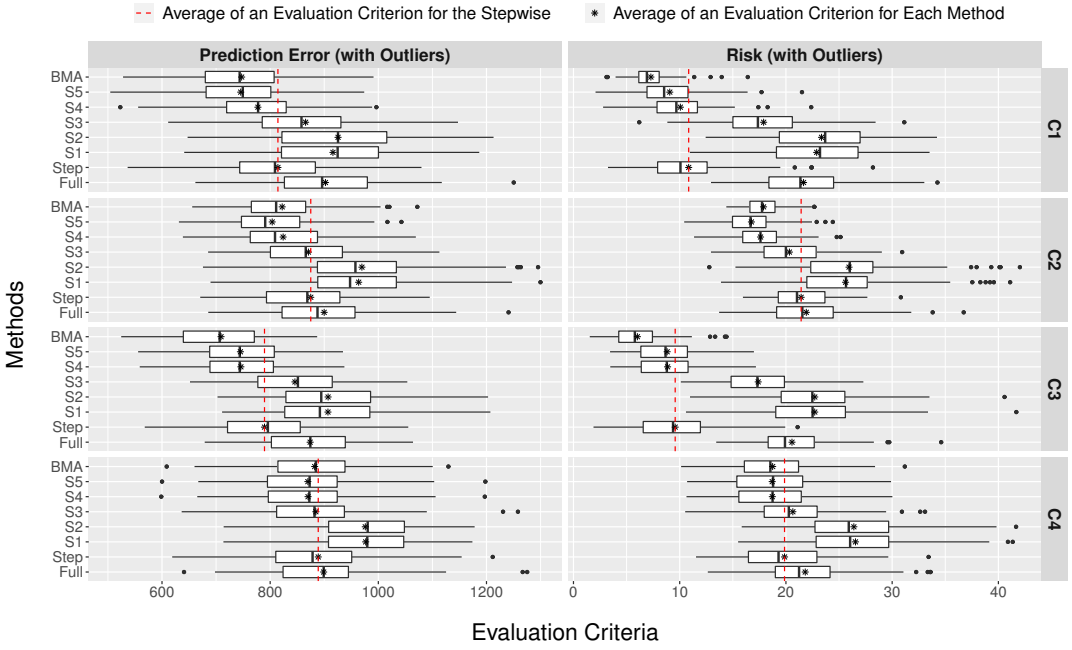

Figure 2: *Boxplots of PEs and Risks over the 100 simulations when a dataset contains many outliers.*

be verified through Figures 1 and 2.

The simulation result agrees with the findings in Breiman (1996b) in the sense that stacking shows better performance when the candidate model list contains dissimilar models, which is more likely for C2 and C4 cases because those coefficient sets contain many significant coefficients. On the other hand, since C1 and C3 contain only three significant coefficients, there are few valid candidate models to give a better performance, and the effect of combining significantly different models in stacking may be minimal.

3.2. When the distributional assumption is violated

When an error term in a linear regression model follows a heavy-tailed distribution rather than a normal distribution (e.g., t-distribution) or a skewed distribution such as a Poisson distribution, it violates the assumption of normality of error terms in the standard linear regression analysis. These error terms may have an effect on the prediction and estimation of parameters if an analysis based on the normality assumption. In particular, if the error terms follow a distribution with heavier tails than the normal distribution, the least squared method can be sensitive to a portion of data (Montgomery *et al.*, 2012). Similarly, for the Poisson response type data, although it would be better to consider an alternative linear model such a generalized linear models (Agresti, 2013), it is possible that we conduct a standard linear regression analysis even with such type of response variable if we don't have enough information of the dataset, and it usually leads to an unreliable results. Thus, in order to deal with such a problem, especially in the prediction and estimation of the coefficients, we may consider the model averaging methods as a technique of the robust regression and compare performance with one from the standard linear regression model.

Table 3: Averages of evaluation criteria over the 100 simulations when the response variable is calculated based on the error term from a heavy-tailed distribution (t_1). The true model is given in (3.5), T1 and T2 represent different sets of coefficients (see (3.6)), and 100 observations are considered for each simulation.

Risks (with errors from t_1)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
T1	36.4001	20.1239	12.6665	25.1513	29.2460	17.9973	**6.7869**	18.9619
(s.e.)	(73.8479)	(30.2004)	(10.6819)	(35.7044)	(57.6830)	(28.2860)	(3.9042)	(30.5946)
T2	19.7780	11.7900	13.0817	18.1046	16.3247	10.1381	**4.8940**	10.5541
(s.e.)	(32.8971)	(18.5006)	(19.4249)	(27.4072)	(24.4871)	(12.9255)	(2.7675)	(13.8909)
Prediction errors (with errors from t_1)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
T1	2575.97	2310.91	1852.60	2372.81	2424.42	2237.97	**1670.07**	2243.53
(s.e.)	(8783.36)	(8642.56)	(8519.11)	(8629.16)	(8712.53)	(8627.20)	(8517.92)	(8620.53)
T2	1498.99	1387.34	1314.32	1500.97	1437.75	1331.72	**1083.46**	1341.52
(s.e.)	(2895.10)	(2842.59)	(2806.94)	(2873.10)	(2852.92)	(2807.36)	(2747.52)	(2810.30)

Suppose that a true linear regression model is expressed by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_8 X_8 + \varepsilon. \tag{3.5}$$

100 samples of each of $(x_{i1}, \ldots, x_{i8})$ are obtained from a standard normal distribution, independently, and two sets of coefficients are considered:

$$\begin{array}{rccccccccc} & \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 & \beta_7 & \beta_8 \\ T1: & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ T2: & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array} \tag{3.6}$$

T1 implies a model with many significant variables, and T2 indicates a model with a few significant variables. The error term $\varepsilon_i,\ i = 1, \ldots, 100$, is generated from a t-distribution with one degree of freedom, and $y_i,\ i = 1, \ldots, 100$, is calculated based on the regression model (3.5). A list of candidate models is also constructed based on the Breiman's approach.

The result of simulation studies over 100 repetitions is given in Table 3, and corresponding boxplots are shown in Figure 3. Note that, for the illustration purpose, we scaled the unit of each metric in Figure 3 with $\log_{10}(\textit{Evaluation Metric})$. According to Table 3, the stacking method under constraints S5 gives the better performance in both risks and prediction errors than the model containing all variables (Full), the model with variables selected by the stepwise method (Step), and BMA. S4 also provides good performance for Risks and PEs. In particular, S5 shows the significant outperformance not only in the sense of the averages of metrics, but also in the sense of the stability.

To evaluate performance of the model averaging methods with Poisson response data, we conduct a similar simulation by generating a dataset based on:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_8 X_{i8}, \quad i = 1, \ldots, 100, \tag{3.7}$$

where $E[Y_i] = \mu_i$ and $Y_i \sim \text{Poisson}(\mu_i)$. Similar to the previous simulation, we use two coefficient sets (3.6) and randomly generated 100 observations $(x_{i1}, \ldots, x_{i8})$ from the standard normal distribution. The response value $y_i,\ i = 1, \ldots, 100$, is a randomly generated based on a Poisson distribution with mean $e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_8 x_{i8}}$. A candidate model list is also chosen based on the Breiman's approach.

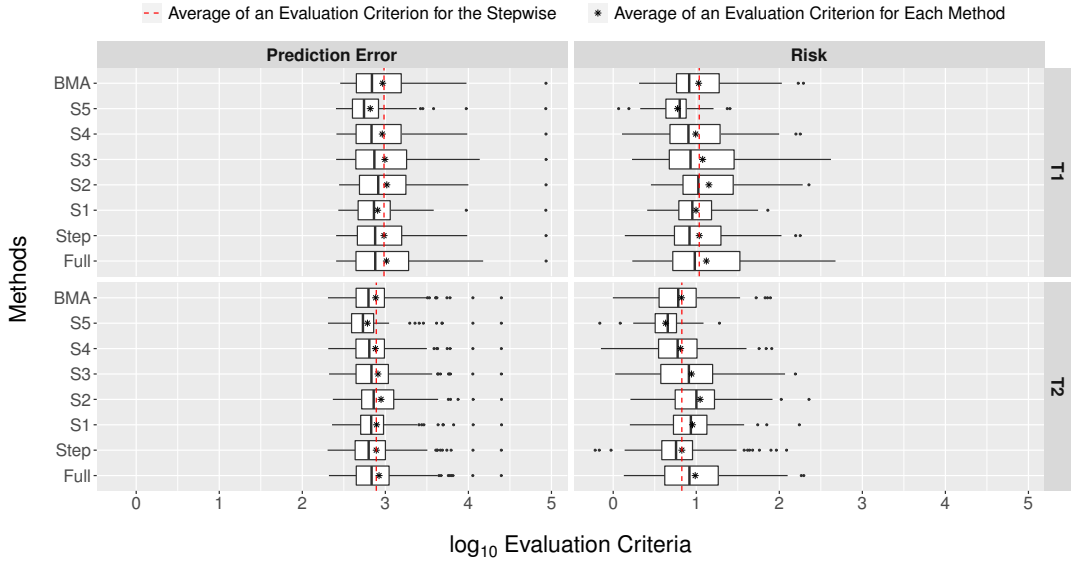

Figure 3: *Boxplots of PEs and Risks over the 100 simulations when the response variable is calculated based on the error term from a heavy-tailed distribution (t_1). Note that evaluation metric units are modified with* $\log_{10}(x)$.

Table 4: Averages of evaluation criteria over the 100 simulations when the response values actually are generated from a Poisson distribution. The true model is given in (3.7), T1 and T2 represent different sets of coefficients (see (3.6)), and 100 observations are considered for each simulation.

Risks (Poisson response data)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
T1	1.7310	1.7439	1.9883	2.0093	1.6865	1.6397	**1.6161**	1.6687
(s.e.)	(0.2199)	(0.1860)	(0.2941)	(0.2889)	(0.1912)	(0.1708)	(0.1714)	(0.1483)
T2	1.7361	1.5797	1.9994	2.0083	1.6871	1.5323	**1.5094**	1.5097
(s.e.)	(0.2079)	(0.1836)	(0.2681)	(0.2650)	(0.1997)	(0.1680)	(0.1679)	(0.1400)
Prediction errors (Poisson response data)								
	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
T1	88.94	89.10	92.59	92.81	88.68	88.43	**88.07**	88.52
(s.e.)	(6.62)	(7.00)	(8.30)	(8.25)	(6.59)	(6.94)	(6.96)	(7.05)
T2	88.85	88.14	91.95	92.08	88.55	87.57	**87.22**	87.24
(s.e.)	(7.58)	(7.27)	(7.53)	(7.53)	(7.49)	(7.17)	(7.23)	(7.12)

Table 4 shows the average of risks and prediction errors over 100 repetitions of the simulation. For both risks and prediction errors, the S5 shows better result compared to Full, Step, and even BMA. Since its performance is pretty similar to the previous simulation, we do not include corresponding boxplots, here.

4. Application to a water pollution data

The effect of land uses on water quality had been given a considerable attention in the 1970s. Haith (1976) collected data to investigate a relationship between land uses and water quality. The dataset

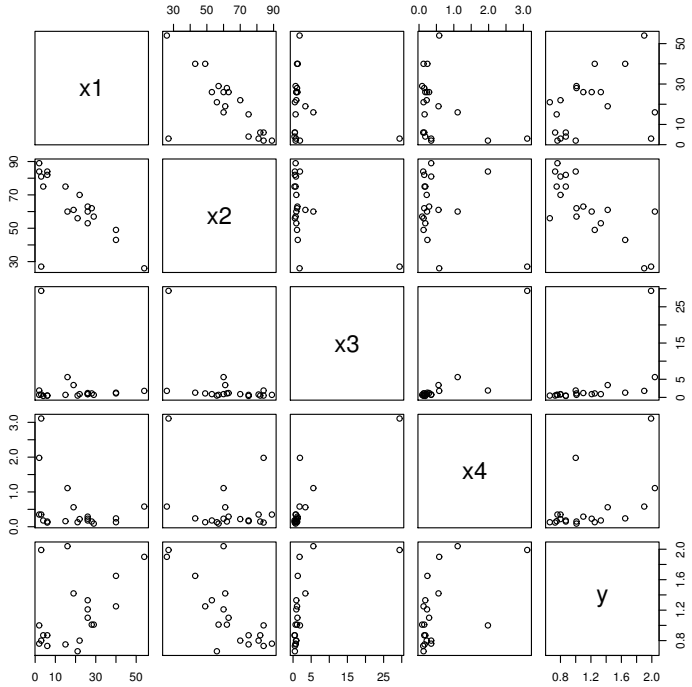

Figure 4: *Scatter plot matrix for the water pollution data.*

consists of observations from twenty river basins in New York State. The response variable Y is the mean concentration of total nitrogen of water samples taken at regular intervals during the spring, summer, and fall months. X_1 indicates the percentage of land area in agricultural use, X_2 is the percentage of land area in forest, forest bushland, and plantations, X_3 implies the percentage of land area in residential use including urban, suburban and rural communities and strip developments with more than four residences per 1,000ft, and X_4 represents the percentage of land area in either commercial or manufacturing use.

The scatterplot matrix in Figure 4 indicates that X_1 and X_2 seem to have a strong negative relationship, and there exist suspected outliers based on the scatterplot for (X_1, X_2) and (X_1, X_3). A linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

is considered to conduct a statistical analysis between land uses and the water quality. The result of the standard linear regression model analysis is shown in Table 5. Even though F-value, 9.15, for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is very significant, all the t-values are small and none of the coefficients are significant. This is a typical indication of existence of multicollinearity among the variables. The *Variance Inflation Factors* (VIF) in Table 5 also supports the multicollinearity problem among X_1, X_2 and X_3 since those VIFs are larger than 10 (Chatterjee and Hadi, 2015).

According to the result in the preliminary analysis, it is likely that one of standard linear regression assumptions is violated, i.e., strong evidence for the multicollinearity, and model averaging methods

Table 5: Summary of standard linear regression analysis with the water pollution data

Variable	Estimates	Standard error	t-value	p-value	VIF
Intercept	1.722	1.234	1.40	0.1832	-
X_1	0.006	0.015	0.39	0.7046	13.28
X_2	−0.013	0.014	−0.93	0.3667	16.73
X_3	−0.007	0.034	−0.21	0.8337	12.68
X_4	0.305	0.164	1.86	0.0823	4.14

$R^2 = 0.71$; adj-$R^2 = 0.63$; F-statistic = 9.15; p-value = 0.0006.

Table 6: Leave-one-out cross-validation estimator of prediction error ($\widehat{\text{PE}}_{\text{LOOCV}}$) over the several methods using the water pollution data

	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
$\widehat{\text{PE}}_{\text{LOOCV}}$	1.5119	1.6324	1.4900	1.4171	1.4186	**1.3921**	1.4112	1.6102

Table 7: Estimated coefficients for each method using the water pollution data

	Full	Step	Stacking					BMA
			S1	S2	S3	S4	S5	
Intercept	1.7222	2.0962	0.9176	1.2627	1.8355	2.2266	2.1616	1.9967
X_1	0.0058	-	0.0097	0.0069	0.0037	0.0005	0.0009	0.0020
X_2	−0.0130	−0.0165	−0.0034	−0.0079	−0.0138	−0.0177	−0.0172	−0.0155
X_3	−0.0072	-	0.0109	0.0051	0.0000	0.0000	0.0000	−0.0004
X_4	0.3050	0.1877	0.8983	0.4758	0.2318	0.0693	0.0723	0.1825

can be an alternative way to improve the performance of analysis in the sense of prediction. Since we cannot generate a new 'test dataset' with the real application data to compare the performance of several fitting methods, *Leave-one-out cross-validation estimator of prediction error* (Friedman *et al.*, 2001) was used as an evaluation metric, which is defined by

$$\widehat{\text{PE}}_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}^{-i}(\mathbf{x}_i)\right]^2, \tag{4.1}$$

where $\hat{f}^{-i}$ indicates the fitted model excluding the i^{th} observation. Table 6 compares $\widehat{\text{PE}}_{\text{LOOCV}}$ for various methods. Note that we do not include the (estimated) standard error of the $\widehat{\text{PE}}_{\text{LOOCV}}$ in Table 6 because the $\widehat{\text{PE}}_{\text{LOOCV}}$ is a summary of prediction errors derived from different environments (leave-one-out based estimators) and, consequently, comparing standard errors is not likely to provide any useful information. Estimated coefficients for each method are also provided in Table 7.

In previous simulation studies, we observe that a model based on stepwise selection method (Step) usually gives the better performance than one from the model based on the all variables (Full). However, according to Table 6, $\widehat{\text{PE}}_{\text{LOOCV}}$ with the water pollution data shows that Full outperforms Step. On the other hand, considering all the methods (including model averaging methods), we see that all the five types of stacking show better results than one based on Full, Step and BMA. Table 7 shows estimated coefficients. It is known that the stepwise method with chosen variables, X_2 and X_4 here, is useful to deal with multicollinearity problem, and, accordingly, the VIF value with the new model ($Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4$) is much less than 10 (1.05392). Since we don't have any knowledge of the *true* coefficients for the real-world dataset, the risk can not be used to evaluate the performance of each method. However, we are able to verify that S4 is also a useful method to select the variables so that it alleviates the multicollinearity. Therefore, we can conclude that not only the stacking method

is useful to obtain better and robust performance in the prediction than one based on traditional linear regression model (including stepwise selection method) or BMA, but also it is helpful for choosing proper variables to ease the multicollinearity problem.

5. Discussion

The model averaging method is useful in the sense of providing a robust performance compared to a traditional linear regression analysis when some of standard regression model assumptions are violated. In this paper, two different scenarios were investigated in the paper to support the assertion. First of all, in order to verify an effect of outliers, whose existence violates one of the standard assumptions about errors in the linear regression, we compare results with a dataset having many outliers and a dataset generated with no outliers. When data do not contain outliers, Bayesian model averaging shows better performance consistently in both risks and prediction errors. On the contrary, when data include many outliers and a true model has many significant predictors, the stacking (S4 and S5) method provides better performance than Bayesian model averaging. Next, two cases are considered to represent a scenario where the distributional assumption is violated in the linear regression model. When error terms are generated from a heavy-tailed distribution (t-distribution with the degree of freedom 1), the stacking (S4 and S5) outperforms Bayesian model averaging. When a response variable is from a skewed distribution like a Poisson distribution, the stacking (S5) gives better performance in almost all cases. In the application to the water pollution data, in which some assumptions in the standard linear regression model such as collinearity are violated, stacking (S4) provides better performance than any other methods in the perspective of the leave-one-out cross-validation estimator of prediction error.

Clarke (2003) argued that the stacking method shows robust performance when a true model is not on the candidate model list. Similarly, in the linear regression model, if some assumptions are violated in the model, the stacking seems to provide the robust result. On the other hand, when the true model is in the candidate models, Bayesian model averaging gives better performance. In the same manner, when the assumptions are satisfied, Bayesian model averaging always outperforms stacking. Our simulation studies show that S4 (non-negative and sum-to-one weights) and S5 (non-negative weights) provide better performances than other restriction types. Therefore, when some of standard linear regression assumptions are violated – for example, when data contains outliers, or a distribution of error terms does not follow normal distribution – S4 or S5 can be used to improve the performance in the perspective of prediction errors and risks. In particular, for the distributional assumption violation, since the stacking selects its combining weights according to differences between response values and predicted values, it will be less sensitive to the data from non-normal distribution, and it seems reasonable that the stacking method works better. However, BMA estimates each coefficient based on the BIC, and it leads to relatively poor results when likelihood functions are not correctly specified. However, the reason that S4 and S5 restrictions give better performances in the stacking method is not yet supported theoretically, so it can be left as an open question.

However, admittedly, the generalized linear model (GLM) will be more ideal model for the Poisson response data in Section 3.1. According to another simulation (which is not included in this paper), GLM (or GLM model constructed by the stepwise selection method, say, GLM-Step) always provides the best performance in terms of risks. Interestingly, for prediction errors, S5 gives slightly better performance than one from GLM (or GLM-Step). Thus, this can be another evidence that the stacking method improves performance in prediction even in such unfavorable conditions. Toward this end, we may consider model averaging methods designed for GLMs to improve performance, but

it is out of scope of the purpose of this paper, so we omit them here. Also note that the performance of model averaging methods could depend on selected candidate models. In this study, since we utilized Breiman's method to construct a list of candidate models, the number of candidate models is the same as the number predictors. It is possible to include more candidate models, but our limited experience indicates that including more number of candidate models did not necessarily guarantee outperformance over the one based on the Breiman's list. One of possible reasons is that the stacking (especially for S4 and S5) and BMA method usually assigns no weights (or near zero weight) to unreliable models regardless of the number of candidate models. As long as the candidate models are well chosen, model averaging would provide a satisfying result even with the small number of candidate models. It would be worthy to conduct additional studies to evaluate the effects of various choices of candidate models. Another possible issue in regression model may occur when the error terms can not be assumed to be independent. When the independent assumption is violated, the relationship among observations can be captured with various types of time series models. If there is clear seasonality, we may use a time-series model directly, but sometimes it may not be obvious, and it has a critical effect on the precision of the OLS estimation. However, the independence assumption is related to the relationship among observations. Different approach other than model averaging may work better since model averaging methods achieve their improvement via combining various models. One possible solution might be to consider not only various candidate models, but also a modified version of the bagging (or boosting). It would be interesting to evaluate the effect of the independent assumption in model averaging methods and propose a model averaging method when this assumption is violated.

Even though we consider only a "raw" risk here, it would be fairer to compare the performance of estimating coefficients with its "standardized" version because some of methods give higher variability of risks across the simulation studies than other methods. In other words, if we can calculate the standard error of estimated coefficients based on the model averaging methods, we divide the raw risk (which is one of the evaluation metrics we consider) by its standard error, and use it as better evaluation criterion. Therefore, it is worthwhile to obtain theoretically derived standard errors in order to investigate performance of various model averaging methods more precisely. Another benefit of having standard errors is that it would allow extending the current analysis with the model averaging method to the hypothesis testing of the estimated coefficients (or weight) with corresponding standard errors. Since most of the estimators of coefficients based on the model averaging method have a complicated form, we also leave this as a future research topic.

References

Agresti A (2013). *Categorical Data Analysis* (3rd ed.), JohnWiley & Sons, NJ.

Ando T and Li KC (2014). A model-averaging approach for high-dimensional regression, *Journal of the American Statistical Association*, **109**, 254–265.

Ando T and Li KC (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models, *The Annals of Statistics*, **45**, 2654–2679.

Breiman L (1996a). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman L (1996b). Stacked regressions, *Machine Learning*, **24**, 49–64.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chatterjee S and Hadi AS (2015). *Regression Analysis by Example*, John Wiley & Sons, New York.

Clarke B (2003). Comparing bayes model averaging and stacking when model approximation error cannot be ignored, *Journal of Machine Learning Research*, **4**, 683–712.

Eklund J and Karlsson S (2007). Forecast combination and model averaging using predictive meansures,

Econometric Reviews, **26**, 329–363.

Fernandez C, Ley E, and Steel MF (2001). Benchmark priors for bayesian model averaging, *Journal of Econometrics*, **100**, 381–427.

Friedman J, Hastie T, and Tibshirani R (2001). *The Elements of Statistical Learning*(Vol. 1), Springer, New York.

Furnival GM and Wilson RW (1974). Regressions by leaps and bounds, *Technometrics*, **16**, 499–511.

Haith DA (1976). Land use and water quality in new york rivers, *Journal of the Environmental Engineering Division* (EEL Proc. Paper 11902), 1–15.

Hjort NL and Claeskens G (2003). Frequentist model average estimators, *Journal of the American Statistical Association*, **98**, 879–899.

Hoeting JA, Madigan D, Raftery AE, and Volinsky CT (1999). Bayesian model averaging: a tutorial, *Statistical Science*, 382–401.

Kass RE and Raftery AE (1995). Bayes factors, *Journal of the American Statistical Association*, **90**, 773–795.

Liang H, Zou G, Wan AT, and Zhang X (2011). Optimal weight choice for frequentist model average estimators, *Journal of the American Statistical Association*, **106**, 1053–1066.

Madigan D and Raftery AE (1994). Model selection and accounting for model uncertainty in graphical models using occam's window, *Journal of the American Statistical Association*, **89**, 1535–1546.

Montgomery DC, Peck EA, and Vining GG (2012). *Introduction to Linear Regression Analysis* (5th ed, Vol. 821), John Wiley & Sons, New York.

Raftery AE (1995). Bayesian model selection in social research, *Sociological Methodology* (Vol. 25), 111–163.

Schapire RE (2003). The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification* (pp. 149–171), Springer.

Wolpert DH (1992). Stacked generalization, *Neural Networks*, **5**, 241–259.

Zhang X, Zou G, Liang H, and Carroll RJ (2020). Parsimonious model averaging with a diverging number of parameters, *Journal of the American Statistical Association*, **115**, 972–984.

Received November 19, 2020; Revised February 5, 2021; Accepted February 12, 2021