Title

stata.com

Intro — Introduction to Bayesian model averaging

Description Remarks and examples References Also see

Description

This entry provides a software-free introduction to Bayesian model averaging (BMA). See [BMA] **BMA commands** for a suite of commands to perform BMA. Also see [BAYES] **Intro** for an introduction to Bayesian analysis.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Brief motivation

Model averaging is a statistical approach that accounts for model uncertainty in your analysis. Instead of relying on just one model, model averaging averages results over multiple plausible models based on the observed data. In Bayesian model averaging (BMA), the "plausibility" of the model is described by the posterior model probability, which is determined using the fundamental Bayesian principles—the Bayes theorem—and applied universally to all data analyses.

Model averaging can be used to account for model uncertainty when estimating model parameters and predicting new observations to avoid overly optimistic conclusions. It is particularly useful in applications with several plausible models, where there is no one definitive reason to choose a particular model over the others. But even if choosing a single model is the end goal, model averaging can be beneficial. For instance, BMA provides a principled way to identify important models and predictors within the considered classes of models. Its framework allows you to learn about interrelations between different predictors in terms of their tendency to appear in a model together, separately, or independently. It can be used to evaluate the sensitivity of the final results to various assumptions about the importance of different models and predictors. And it provides optimal predictions in the log-score sense.

What is model averaging and why do we need it?

The concept of a model is central in statistics. Classical statistical inference is based on the assumption that there is an underlying data-generating model (DGM), and we can infer its characteristics from the observed data. Selecting an appropriate model for the problem at hand is the first and crucial step in performing statistical analyses. In some applications, we may have a strong theoretical or empirical evidence about the DGM. In other applications, usually of complex and unstable nature, such as those in economics, psychology, and epidemiology, choosing a single reliable model can be difficult. In such cases, it is important to have a principled way to account for the uncertainty in the model-selection process. In practice, we are often interested in a particular property or quantity of the DGM. The classical inferential approach involves choosing a model and estimating this quantity from the observed data conditional on the chosen model. One drawback of working with a single model is that we may assign more precision to our estimates than is supported by the data (Chatfield 1995 and Draper 1995). In predictive inference, single-model approaches do not utilize all available information and may be unstable; see, for example, Piironen and Vehtari (2017).

The model averaging approach is conceptually different. Instead of choosing one model, we consider a list of candidate models. The quantity of interest is then estimated by an average across individual model estimates. Averaging is weighed by how likely each model is. In this way, model averaging accounts for the model-selection uncertainty.

The true DGM may or may not be in our list of candidate models. If it is, classical model-selection approaches may work well. Otherwise, the larger the candidate model space is, the greater the possibility of model selection to choose an incorrect model and make wrong conclusions. And the selected model may change every time the new data become available. In model selection, it may not be clear what constitutes a good candidate, given that the true model is unknown. Popular information-based criteria such as the Bayesian information criterion measure how well a model fits the data and include an additional penalty for its complexity. But a model often needs to be evaluated based on its predictive performance. Improving predictive performance motivated a variety of methods known as ensemble methods such as stacking (Wolpert 1992) and bagging (Breiman 1996). Model averaging can be viewed as an ensemble method that improves predictive performance using optimal combinations in the space of considered candidate models (Raftery and Zheng 2003).

Bayesian model averaging (BMA)

BMA (Leamer 1978) casts model averaging into a Bayesian framework. It provides a principled way to define model weights as posterior model probabilities, which is universal to all data-generating processes. BMA formulation emerges naturally as an application of a standard Bayesian predictive approach to model averaging.

In BMA, model M is a random variable with prior $P(M)$ distributed over some model space. Given the observed data D, the likelihood of M is the probability of D with respect to M, $P(D|M)$. The posterior of M is then given by the Bayes theorem

$$P(M|D) = \frac{P(D|M)P(M)}{\sum_{M^*} P(D|M^*)P(M^*)}$$

where we assume that the model space is discrete and take the sum over it in the denominator. Continuous model spaces are also possible but will not be considered here. The quantity $P(D|M)$ is known as the marginal likelihood of model M. And $P(M|D)$ is known as the posterior model probability and is a key quantity in BMA inference and prediction. Also see *Concepts of BMA*.

Let Q be any quantity of interest that is not model specific; that is, it should have the same interpretation across all models in the model space. Let Q_M be its estimator with respect to model M. The BMA estimator of Q can be written as

$$Q_{\text{BMA}} = \sum_M P(M|D) Q_M$$

The above formula follows from the fundamental BMA formula for the posterior distribution of Q over the model space,

$$g(Q|D) = \sum_M P(M|D) g(Q|D, M)$$

where $g(Q|D, M)$ is the posterior distribution of Q for model M. Then, $Q_{\text{BMA}} = E(Q|D)$ is the posterior mean of Q, and $Q_M = E(Q|D, M)$ is the posterior mean of Q for model M.

The variability of Q is described by the posterior variance of Q with respect to $g(Q|D)$,

$$\text{Var}(Q|D) = \sum_M P(M|D) \text{Var}(Q|D, M) + \sum_M P(M|D) \{E(Q|D, M) - E(Q|D)\}^2$$

where the second term estimates the additional uncertainty about the estimated Q across models.

In a regression context, the notion of model averaging has a more specific formulation—model uncertainty arises mainly from the uncertainty of which predictors should be included in the model.

Let Y be an outcome variable with p potential predictors (regressors or covariates) $\mathbf{x} = (X_1, X_2, \ldots, X_p)$. Let $D = \{y_i, x_{1i}, x_{2i}, \ldots, x_{pi}\}_{i=1}^n$ be a sample of observations on Y and $\mathbf{x}$. We are not sure which predictors describe Y best and consider any subset of $\mathbf{x}$ as a potential candidate set. We can enumerate all subsets and denote the jth subset by $\mathbf{x}_j$. Then M_j, defined as the model corresponding to $\mathbf{x}_j$, is an element of the discrete model space $\{M_j\}_{j=1}^{2^p}$. Two typical applications of BMA in this context are estimating regression coefficients and predicting Y from a new observation $\mathbf{x}^*$; also see *Usage of BMA* for other applications.

Let $\widehat{\boldsymbol{\beta}}_{M_j}$ be an estimate of a $p \times 1$ regression coefficient vector $\boldsymbol{\beta}$ with respect to model M_j, in which the coefficients for predictors not in the model are set to zero. Then, the BMA estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{BMA}} = \sum_{j=1}^{2^p} P(M_j|D) \widehat{\boldsymbol{\beta}}_{M_j}$$

Given a new observation $\mathbf{x}^*$, a new outcome value y^* can be obtained from the BMA predictive distribution, which is as a mixture of the model-specific predictive distributions,

$$p_{\text{BMA}}(y^*|\mathbf{x}^*, D) = \sum_{j=1}^{2^p} P(M_j|D) p_j(y^*|\mathbf{x}^*, D, M_j)$$

where $p_j(y^*|\mathbf{x}^*, D, M_j)$ is the posterior predictive density of Y for model M_j. The above is a special case of the standard definition of the Bayesian posterior predictive distribution.

BMA has many appealing statistical properties, as detailed in Steel (2020). For instance, Raftery and Zheng (2003) show that BMA point estimators and predictors minimize the mean squared error weighted by the prior. Madigan and York (1995) verify that BMA estimators maximize predictive ability, and Min and Zellner (1993) find that BMA performs better than any other model-choice approach with respect to the log predictive-score.

BMA is built on Bayesian principles. Thus, it faces the same challenges as standard Bayesian analyses with respect to the specification of priors and the intensity of the computations. We discuss some of these challenges in more detail in the next section.

Concepts of BMA

Here we briefly describe some of the concepts essential to BMA. For details in the context of linear regression, see *Remarks and examples* and *Methods and formulas* of [BMA] **bmaregress**. Also see Hoeting et al. (1999), Fernández, Ley, and Steel (2001a), Moral-Benito (2015), Fragoso, Bertoli, and Louzada (2018), and Steel (2020).

BMA applies standard Bayesian principles to model averaging. Thus, all concepts of Bayesian analysis apply to BMA as well; see [BAYES] **Intro**. Compared with standard Bayesian analyses, which condition on a model, BMA views a model as random and assumes a prior distribution for it.

Model space. The extent to which BMA can properly account for model uncertainty relies on the construction of the model space. BMA results are conditional on the considered model space. If the model space does not include important candidate models, BMA will not be able to consider them and incorporate them in the results. The model space should incorporate any aspects of model uncertainty that needs to be accounted for. For instance, if one is uncertain about various functional forms of predictors, these functional forms (and possibly more) should be included in the model space. Ideally, the model space should contain the DGM, but BMA was found to provide good results even when it does not, as long as the model space is sufficiently large. In that case, BMA approximates the true DGM by a combination of models within the considered class. See Steel (2020) for a detailed discussion of the construction of the BMA model space.

Parameters of interest. When the goal of analysis is an estimation of a parameter of interest, it is important that the parameter has the same interpretation across all models. For instance, see *Interpretation of BMA regression coefficients* in *Remarks and examples* of [BMA] **bmaregress**.

Priors for models and model parameters. Specifying a prior distribution for a model parameter is an integral part of a Bayesian model specification. BMA additionally specifies a prior distribution for a model, typically, over a discrete model space. A variety of model priors and priors for model parameters are suggested in the literature, both informative and noninformative, data agnostic and data driven (Steel 2020).

In the regression context, commonly used priors, such as a Zellner's prior with a fixed g parameter for regression coefficients (Fernández, Ley, and Steel 2001a), provide exact computation of marginal likelihoods. Although computationally convenient, these priors may not always provide the best predictive performance. The application of g-priors with random g parameters (Ley and Steel 2012) allows for more flexible BMA analysis but complicates the model specification and simulation. See *Introduction to BMA for linear regression* in *Remarks and examples* of [BMA] **bmaregress** for the discussion of various priors in the context of BMA linear regression.

As with any Bayesian analysis, in the absence of strong information in the data about the DGM and model parameters, BMA results can be sensitive to the choice of priors. Sensitivity analysis is recommended to investigate the impact of priors on the results.

Estimation: Model enumeration and Markov chain Monte Carlo (MCMC) sampling. Depending on the model complexity, it may be feasible to enumerate and consider all the models in the defined space. In this case, the model space is fully explored. This is rarely feasible in practice. More commonly, MCMC sampling is used to explore the model space more efficiently by considering only more likely models given the observed data, for example, the MCMC model composition (MC3) sampling proposed by Madigan and York (1995). In addition to sampling of the model space, we may also need to use MCMC sampling for model parameters when analytical expressions for their posterior distributions are not available, which is common in practice. When MCMC sampling is used, it is important to verify the convergence of MCMC; see *Convergence of BMA* in *Remarks and examples* of [BMA] **bmaregress** in the context of linear regression.

Posterior model probability (PMP). The PMP is central to all BMA analyses. It represents the probability of a model given the observed data and model's prior. It is used as a weight in BMA estimates of parameters of interest and predictions. It is used to identify influential models. And it is used to compute the posterior inclusion probability (PIP), which is used to identify important predictors. In special cases, the PMP can be estimated exactly or analytically, in which case we refer to it as the analytical PMP. More commonly, however, it is estimated based on the MCMC sample of models, in which case we refer to it as the frequency PMP. Models with high PMPs are of interest in BMA analysis.

Posterior inclusion probability (PIP). The PIP is the probability that a predictor is included in a model computed over the model space given the observed data and the prior model probability. It measures the importance of a predictor. Because the computation of the PIP is based on the PMP, we also distinguish between the analytical PIP and frequency PIP. Predictors with high PIP values, commonly above 0.5, are considered important predictors.

Jointness. Jointness is a concept particular to BMA. Because BMA considers multiple models, it can estimate the tendency of predictors to be included jointly or exclusively across the models. Jointness means that predictors tend to be included together in many models. Such predictors are then viewed as complements, in the sense that their joint inclusion provides additional information in explaining the outcome. Disjointness means that whenever one predictor is included in a model, the other tends to be excluded. Such predictors are viewed as substitutes, meaning that only one of them is needed to explain the outcome.

Inference. In the context of BMA, the inference focuses on exploring influential models, models with high PMPs, and important predictors, predictors with high PIPs. The jointness or disjointness of predictors is often also of interest. When averaging across the model space is applicable for a parameter of interest, the parameter estimation is performed with respect to the posterior distribution over the model space. Although the inference accounts for model uncertainty, it is important to remember that it is still conditional on the explored model space.

Prediction. BMA is commonly used for prediction because of its theoretical properties and empirical performance. When the model space contains the DGM, the BMA predictive mean minimizes the expected squared error loss (Min and Zellner 1993). Madigan and Raftery (1994) compare the BMA predictive performance with that of a single model using the log predictive-score (LPS) and conclude that BMA performs at least as well. See Steel (2020) for more information.

Log predictive-score (LPS). LPS is the negative of the logarithm of the predictive density evaluated at an observation (Good 1952). It is used to assess predictive performance of a model in the context of BMA (for example, Madigan, Gavrin, and Raftery [1995] and Fernández, Ley, and Steel [2001a]). It can also be used to compare model fit.

Diagnostics. Model diagnostics are just as important for BMA as they are for a single-model analysis. Any model checks that are commonly done with one model should be performed during BMA analysis as well. Because of the many models, the application of such checks is not as straightforward. The

literature recommends that the checks be performed for the model with all predictors before the estimation and for all high-PMP models after the estimation. For BMA, additional diagnostics include checking MCMC convergence and performing a sensitivity analysis to the prior choices.

Sensitivity analysis. As with any Bayesian analysis, prior sensitivity analysis is important for BMA. In the BMA context, the sensitivity analysis should be performed for both model priors and model parameter priors.

Usage of BMA

Fragoso, Bertoli, and Louzada (2018) identified several main applications of BMA across various disciplines such as "model choice", "combination of multiple models for prediction", and "combined estimation". We will refer to these simply as model choice, parameter estimation, and prediction.

BMA was motivated in the context of prediction to improve out-of-sample predictive performance of a model (for example, Hoeting et al. [1999]). BMA can be shown to produce optimal predictions with respect to the LPS (Min and Zellner 1993) by averaging predictions from multiple models and weighing them by the model's importance. The model's importance is estimated in a principled Bayesian way as a PMP and applied universally to all data-generating processes. A few applications of BMA for prediction can be found in Madigan and Raftery (1994), Raftery, Madigan, and Volinsky (1995), Volinsky et al. (1997), Hoeting et al. (1999), Tobias and Li (2004), Kaplan and Lee (2018), and Darwen (2019).

The use of BMA for model choice amounts to identifying important models and predictors. The importance of a model is based on the estimated PMP. And the importance of a predictor is based on the estimated PIP, the probability that this predictor is included in a model estimated over the considered model space. Some of the applications of model choice include Raftery, Madigan, and Hoeting (1997), Hoeting et al. (1999), Fernández, Ley, and Steel (2001b), Eicher, Papageorgiou, and Raftery (2011), Moral-Benito (2015), Arin and Braunfels (2018), and Peisker (2023).

BMA is also used to estimate a parameter common to all models. As with prediction, the BMA estimate is a weighted average of the model-specific estimates with weights defined by PMPs. For instance, see Hoeting et al. (1999), Koop (2003), Yin and Yuan (2009), Montgomery and Nyhan (2010), and Moral-Benito (2015). But be mindful when using BMA to estimate partial regression coefficients in a linear regression (Draper 1999; Banner and Higgs 2017); see *Interpretation of BMA regression coefficients* in *Remarks and examples* of [BMA] **bmaregress**.

Wasserman (2000) also shows how to use BMA to perform Bayesian variable selection.

See Fragoso, Bertoli, and Louzada (2018) for more references and discussion of the BMA usage in different research areas.

BMA versus frequentist model averaging

Frequentist model averaging (FMA) is an inferential procedure based on the so-called FMA estimator,

$$\widehat{\boldsymbol{\beta}}_{\text{FMA}} = \sum_{j=1}^{2^p} \omega_j \widehat{\boldsymbol{\beta}}_j$$

where $0 \leq \omega_j \leq 1$, $\sum_{j=1}^{2^p} \omega_j = 1$, and $\widehat{\boldsymbol{\beta}}_j$ is an estimator, usually ordinary least squares, of regression parameters for model M_j. The weights ω_j's are chosen such that $\widehat{\boldsymbol{\beta}}_{\text{FMA}}$ has certain asymptotic properties.

In contrast to BMA, where model estimators are weighted by PMPs, in FMA the weights are computed for each model independently and then normalized. The most common choice is $\omega_j \propto \exp(-0.5I_j)$ (Buckland, Burnham, and Augustin 1997), where I_j is an information criterion of the form

$$I_j = -2\log(\widehat{L}_j) + \psi_j$$

This approach includes popular choices such as the Akaike information criterion, $\psi_j = 2p_j$, and Bayesian information criterion, $\psi_j = p_j \log(n)$, where p_j is the number of predictors in the jth model. Other approaches include weights based on Mallow's criterion (Hansen 2007) and cross-validation (Hansen and Racine 2012). A more in-depth exploration of the FMA, as applied in economics in particular, can be found in Moral-Benito (2015).

Compared with FMA, BMA provides a unified and intuitive way to interpret the model's and predictor's importance by using the respective PMPs and PIPs. In fact, the PMPs, which are derived from fundamental Bayesian principles, are used as weights in all BMA computations. BMA can also handle larger model spaces more easily by using efficient MCMC sampling algorithms. Additionally, BMA benefits from several appealing statistical properties such as calibration of credible intervals and optimal prediction in the log-score sense. See Steel (2020) for details.

See De Luca and Magnus (2011) for the implementation of the weighted-average least-squares estimator in Stata.

Computational methods for BMA

For a long time, the use of BMA in practice has been hindered by the lack of computationally feasible estimation methods. Since then, a variety of specialized MCMC methods have been developed to facilitate Bayesian inference. A unique challenge of BMA is the complex nature of the posterior domain—a discrete mixture of models with continuous domains of varying dimensions.

One of the first general sampling methods for BMA was the MC3 (Madigan and York 1995), which is a stochastic method that moves through the model space by changing one predictor, or a group of predictors, at a time.

The availability of the analytical form for the marginal likelihood in linear models leads to fast and efficient MC3 sampling methods. However, analytical marginals are not available for generalized linear models and for most linear BMA models that include hyperparameters such as g-priors. Ley and Steel (2012) proposed an adaptive MC3 method applicable to the latter case. Other adaptive MCMC methods are also available (Atchadé and Rosenthal 2005).

Motivating examples

Consider the following simulated dataset. There are $n = 200$ observations and $p = 10$ predictors. Each predictor `x1` through `x10` is generated independently from a standard normal distribution. The outcome `y` is generated according to the following regression model, which we refer to as our DGM,

$$\texttt{y} = 0.5 + 1.2 \times \texttt{x2} + 5 \times \texttt{x10} + \epsilon$$

where $\epsilon \sim N(0, 1)$ is a standard normal error term.

```
. use https://www.stata-press.com/data/r18/bmaintro
(Simulated data for BMA example)
. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
y	200	.9944997	4.925052	-13.332	13.06587
x1	200	-.0187403	.9908957	-3.217909	2.606215
x2	200	-.0159491	1.098724	-2.999594	2.566395
x3	200	.080607	1.007036	-3.016552	3.020441
x4	200	.0324701	1.004683	-2.410378	2.391406
x5	200	-.0821737	.9866885	-2.543018	2.133524
x6	200	.0232265	1.006167	-2.567606	3.840835
x7	200	-.1121034	.9450883	-3.213471	1.885638
x8	200	-.0668903	.9713769	-2.871328	2.808912
x9	200	-.1629013	.9550258	-2.647837	2.472586
x10	200	.083902	.8905923	-2.660675	2.275681

We consider three toy examples. The first example briefly introduces BMA for linear regression and compares it with standard linear regression. The second example compares the use of regression, stepwise selection, lasso, and BMA for prediction. The third example revisits these tools in a more challenging setting of $n = p$.

Examples are presented under the following headings:

▷ Example 1: BMA linear regression

We first use `regress` to fit a standard linear regression of `y` on `x1` through `x10`. We specify the predictors by using the shortcut *varlist* notation `x1-x10`:

```
. regress y x1-x10
```

Source	SS	df	MS		
Model	4607.24837	10	460.724837	Number of obs	= 200
Residual	219.723235	189	1.1625568	F(10, 189)	= 396.30
				Prob > F	= 0.0000
Total	4826.9716	199	24.2561387	R-squared	= 0.9545
				Adj R-squared	= 0.9521
				Root MSE	= 1.0782

y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x1	.0753537	.0781737	0.96	0.336	-.0788513	.2295587
x2	1.18854	.0716658	16.58	0.000	1.047172	1.329907
x3	-.1871012	.0789484	-2.37	0.019	-.3428344	-.0313679
x4	-.0459335	.0785503	-0.58	0.559	-.2008813	.1090144
x5	.0343498	.0793095	0.43	0.665	-.1220956	.1907953
x6	-.0149194	.0767357	-0.19	0.846	-.1662879	.136449
x7	.007174	.0831239	0.09	0.931	-.1567958	.1711437
x8	-.0384917	.0810626	-0.47	0.635	-.1983953	.1214119
x9	.0968948	.0817218	1.19	0.237	-.0643093	.2580989
x10	5.13251	.0877447	58.49	0.000	4.959426	5.305595
_cons	.617996	.0791152	7.81	0.000	.4619337	.7740582

`regress` identifies the two true predictors `x2` and `x10` as "statistically significant" (with p-values less than 0.000). The estimate of the coefficient for `x2` is 1.19 with a standard error of 0.072, and the

95% confidence interval (CI) is $[1.05, 1.33]$, which agrees with the true value of 1.2. The estimated coefficient for `x10` is 5.13 with a standard error of 0.088, and the 95% CI is $[4.96, 5.31]$, which agrees with the true value of 5. These findings are consistent with our true DGM. `regress` also reports a p-value of 0.019 for `x3`, which is not in the DGM, with an estimated coefficient of -0.19 and a 95% CI of $[-0.34, -0.03]$. It might be tempting to use the reported p-values to infer the importance of the predictors, but p-values do not have such interpretation.

Let's now use `bmaregress` to perform BMA for a linear regression:

```
. bmaregress y x1-x10

Enumerating models ...
Computing model probabilities ...

Bayesian model averaging                        No. of obs          =    200
Linear regression                               No. of predictors   =     10
Model enumeration                                         Groups    =     10
                                                          Always    =      0
Priors:                                         No. of models       =  1,024
  Models: Beta-binomial(1, 1)                       For CPMP >= .9  =      9
   Cons.: Noninformative                        Mean model size     =  2.479
   Coef.: Zellner's g
       g: Benchmark, g = 200                    Shrinkage, g/(1+g)  = 0.9950
  sigma2: Noninformative                        Mean sigma2         =  1.272
```

y	Mean	Std. dev.	Group	PIP
x2	1.198105	.0733478	2	1
x10	5.08343	.0900953	10	1
x3	-.0352493	.0773309	3	.21123
x9	.004321	.0265725	9	.051516
x1	.0033937	.0232163	1	.046909
x4	-.0020407	.0188504	4	.039267
x5	.0005972	.0152443	5	.033015
x8	-.0005639	.0153214	8	.032742
x7	-8.23e-06	.015497	7	.032386
x6	-.0003648	.0143983	6	.032361
Always				
_cons	.5907923	.0804774	0	1

```
Note: Coefficient posterior means and std. dev. estimated from 1,024 models.
Note: Default priors are used for models and parameter g.
```

We will describe only some of the more relevant information here, but see example 1 of [BMA] **bmaregress** for details about the output of `bmaregress`.

`bmaregress`, with the default settings, considered all $2^{10} = 1{,}024$ possible models based on 10 predictors. Like `regress`, `bmaregress` identified the two true predictors, `x2` and `x10`, with the estimated PIPs of 1, labeled as `PIP` in the table. All other predictors have much lower `PIP` values, and all but the `PIP` for `x3` are below 10%. Unlike `regress`, we can use the `PIP` reported by `bmaregress` to describe and compare the importance of predictors. `PIP` genuinely represents the probability of a predictor being included in a model across the considered space of 1,024 possible models. For instance, the `PIP` of 0.2 for `x3` is much lower than that for `x2` and `x10`, so we can conclude that this predictor is not as important. Also, its BMA coefficient (posterior mean) of -0.035 is much closer to 0 than that from `regress`.

The BMA estimates of 1.2 (rounded) and 5.1 of the coefficients for `x2` and `x10`, respectively, are close to the true values of 1.2 and 5. The respective estimated posterior standard deviations, 0.073 and 0.090, are slightly larger than those from `regress`. This is expected because the BMA estimates account for the uncertainty about which predictors should be included in the regression

model. bmaregress does not report credible intervals by default for computational reasons, but you can obtain them as described in example 5 of [BMA] **bmaregress**. Also, with real-world observational data, we should be mindful when interpreting BMA regression coefficients; see *Interpretation of BMA regression coefficients* in *Remarks and examples* of [BMA] **bmaregress**.

Although BMA does not "select" a model, it does identify some of the influential models that contribute more to the averaged results. In this example, we can already guess which model BMA identified as the top model based on the reported PIP values by bmaregress, but let's use bmastats models to confirm:

```
. bmastats models
Computing model probabilities ...
Model summary          Number of models:
                                Visited = 1,024
                               Reported =     5
```

	Analytical PMP	Model size
Rank		
1	.6292	2
2	.1444	3
3	.0258	3
4	.0246	3
5	.01996	3

Variable-inclusion summary

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
x2	x	x	x	x	x
x10	x	x	x	x	x
x3		x			
x9			x		
x1				x	
x4					x

Legend:
x - estimated

As anticipated, the top model with a PMP, Analytical PMP, of 0.63 is the model that contains x2 and x10. The next plausible model based on our sample is the one that also includes x3, but its PMP of 0.14 is much lower.

In the above, bmaregress used the default priors. These priors are offered for convenience and should be carefully evaluated in each application. Also, sensitivity analysis should be performed to evaluate the impact of different priors on the results; see, for example, example 11 of [BMA] **bmaregress**.

In BMA, the variance of the prior for the regression coefficients is proportional to the so-called g parameter. By default, g has a fixed value of $\max(n, p^2)$, which in our example is $g = n = 200$. We can relax this by specifying a higher value for g, say, 1,000. This will reduce the shrinkage effect on the coefficients and generally produce estimates that are closer to the ordinary least-squares estimates.

Another important benefit of BMA is its ability to control model uncertainty through the model prior. If, for example, we had a prior knowledge that predictors x1 and x3 through x9 are unlikely to be related to y, we could incorporate this knowledge in our BMA model. In the following specification, we use the mprior() option to specify the binomial model prior with the inclusion probability of 0.1 for x1 and x3 through x9 and the inclusion probability of 0.5 for x2 and x10.

```
. bmaregress y x1-x10, mprior(binomial x2 x10 0.5 x1 x3-x9 0.1)
Enumerating models ...
Computing model probabilities ...
Bayesian model averaging                        No. of obs         =    200
Linear regression                               No. of predictors  =     10
Model enumeration                                         Groups   =     10
                                                          Always   =      0
Priors:                                         No. of models      =  1,024
  Models: Binomial, IP varies                       For CPMP >= .9 =      2
   Cons.: Noninformative                        Mean model size    =  2.129
   Coef.: Zellner's g
       g: Benchmark, g = 200                    Shrinkage, g/(1+g) = 0.9950
  sigma2: Noninformative                        Mean sigma2        =  1.276
```

y	Mean	Std. dev.	Group	PIP
x2	1.200944	.0730381	2	1
x10	5.080663	.0899736	10	1
x3	-.0106068	.0452704	3	.064039
x9	.0009677	.0126993	9	.012195
x1	.0008208	.0115323	1	.01149
Always				
_cons	.5884159	.0803504	0	1

```
Note: Coefficient posterior means and std. dev. estimated from 1,024 models.
Note: Default prior is used for parameter g.
Note: 5 predictors with PIP less than .01 not shown.
```

The effect of this model prior is that the posterior inclusion probability of predictors `x1` and `x3` through `x9` is now less than 8%. There is also a slight improvement in the estimates of the intercept and the coefficient for `x2`.

The inclusion of prior assumptions supported by science and empirical work in a model is part of standard Bayesian analysis. With such priors, the BMA framework has the potential to provide a more reliable inference than the classical regression approach in the situations where the data have limited information about the model and its parameters.

◁

▷ Example 2: BMA for prediction compared with other approaches

In this example, we compute and compare predictions for the `bmaintro` dataset by using the following methods: linear regression, `regress` (see [R] **regress**); stepwise selection with linear regression, the `stepwise` prefix (see [R] **stepwise**); linear lasso variable selection, `lasso linear` (see [LASSO] **lasso**); and BMA linear regression, `bmaregress` (see [BMA] **bmaregress**).

To compare predictive performance of the models, we split our dataset into two equal samples: one for "training" the model (used for fitting) and the other for "testing" the model (used for prediction). We store the resulting sample identifier in the `sample` variable. And we specify a random-number seed for reproducibility.

```
. splitsample, generate(sample) nsplit(2) rseed(50)
```

Next, we fit each of the four commands using the training data, `if sample == 1`, and compute predictions using the test data, `if sample == 2`.

We start with `regress` to fit a linear regression and `predict` to obtain the linear predictor for `y`, which we store in the `yreg` variable.

```
. regress y x1-x10 if sample == 1
```

Source	SS	df	MS	Number of obs	=	100
				F(10, 89)	=	199.77
Model	2353.4317	10	235.34317	Prob > F	=	0.0000
Residual	104.84695	89	1.17805562	R-squared	=	0.9573
				Adj R-squared	=	0.9526
Total	2458.27865	99	24.8310975	Root MSE	=	1.0854

y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x1	.2278093	.1115858	2.04	0.044	.0060906	.449528
x2	1.040423	.1084559	9.59	0.000	.824924	1.255923
x3	-.2557993	.1140321	-2.24	0.027	-.4823787	-.0292199
x4	-.0182061	.1175268	-0.15	0.877	-.2517293	.2153171
x5	.0389276	.1187846	0.33	0.744	-.1970948	.27495
x6	.0120724	.1107333	0.11	0.913	-.2079523	.2320971
x7	.0792028	.1378848	0.57	0.567	-.1947713	.3531768
x8	-.0841665	.1259057	-0.67	0.506	-.3343384	.1660054
x9	.0039031	.1181302	0.03	0.974	-.2308191	.2386254
x10	5.281029	.1298317	40.68	0.000	5.023056	5.539002
_cons	.5726978	.1175907	4.87	0.000	.3390475	.8063481

```
. predict yreg if sample == 2
(option xb assumed; fitted values)
(100 missing values generated)
```

Next, we use `stepwise` to perform stepwise backward selection with the significance level of 0.05 for the removal of a predictor from the model. And we use `predict` to obtain the linear predictor from the selected model and store it in the `ysw` variable.

```
. stepwise, pr(.05): regress y x1-x10 if sample == 1
Wald test, begin with full model:
p = 0.9737 >= 0.0500, removing x9
p = 0.9134 >= 0.0500, removing x6
p = 0.8862 >= 0.0500, removing x4
p = 0.7456 >= 0.0500, removing x5
p = 0.5099 >= 0.0500, removing x8
p = 0.5102 >= 0.0500, removing x7
```

Source	SS	df	MS	Number of obs	=	100
				F(4, 95)	=	527.09
Model	2352.28746	4	588.071866	Prob > F	=	0.0000
Residual	105.991183	95	1.11569666	R-squared	=	0.9569
				Adj R-squared	=	0.9551
Total	2458.27865	99	24.8310975	Root MSE	=	1.0563

y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x1	.2143505	.1062623	2.02	0.046	.0033932	.4253078
x2	1.038816	.1028192	10.10	0.000	.8346945	1.242938
x3	-.2465552	.1087814	-2.27	0.026	-.4625137	-.0305968
x10	5.285204	.1231611	42.91	0.000	5.040698	5.52971
_cons	.5527609	.1064491	5.19	0.000	.3414327	.7640891

```
. predict ysw if sample == 2
(option xb assumed; fitted values)
(100 missing values generated)
```

We then use lasso for the linear model followed by lassocoef to see the coefficient estimates from the selected model and by predict to compute and store the penalized linear predictor in the ylasso variable.

```
. lasso linear y x1-x10 if sample == 1, rseed(18) nolog
Lasso linear model                          No. of obs        =        100
                                            No. of covariates =         10
Selection: Cross-validation                 No. of CV folds   =         10
```

ID	Description	lambda	No. of nonzero coef.	Out-of-sample R-squared	CV mean prediction error
1	first lambda	4.697569	0	-0.0072	24.75885
43	lambda before	.0943851	4	0.9506	1.213251
* 44	selected lambda	.0860002	4	0.9507	1.211054
45	lambda after	.0783602	4	0.9507	1.211522
48	last lambda	.0592766	5	0.9503	1.220737

```
* lambda selected by cross-validation.
. lassocoef, display(coef) nolegend
```

	active
x1	.1167579
x2	1.051272
x3	-.1659852
x10	4.53756
_cons	0

```
. predict ylasso if sample == 2
(options xb penalized assumed; linear prediction with penalized coefficients)
```

Finally, we use bmaregress to fit a BMA linear regression followed by bmapredict to compute the posterior predictive mean and store it in the ybma variable.

```
. bmaregress y x1-x10 if sample == 1
Enumerating models ...
Computing model probabilities ...
Bayesian model averaging                        No. of obs         =    100
Linear regression                               No. of predictors  =     10
Model enumeration                                         Groups   =     10
                                                          Always   =      0
Priors:                                         No. of models      =  1,024
  Models: Beta-binomial(1, 1)                       For CPMP >= .9 =     17
   Cons.: Noninformative                        Mean model size    =  2.804
   Coef.: Zellner's g
       g: Benchmark, g = 100                    Shrinkage, g/(1+g) = 0.9901
  sigma2: Noninformative                        Mean sigma2        =  1.412
```

y	Mean	Std. dev.	Group	PIP
x10	5.18159	.1381456	10	1
x2	1.068169	.115504	2	1
x3	-.0676021	.1264303	3	.27554
x1	.0439351	.1014456	1	.20554
x8	-.0043739	.0369172	8	.05923
x9	-.0020804	.0305354	9	.054026
x7	.0022291	.0354666	7	.053837
x5	.0017863	.0301671	5	.053101
x6	.0004583	.0266441	6	.051342
x4	-.0000354	.0281472	4	.051285
Always				
_cons	.5575281	.1202808	0	1

```
Note: Coefficient posterior means and std. dev. estimated from 1,024 models.
Note: Default priors are used for models and parameter g.
. bmapredict ybma if sample == 2, mean
note: computing analytical posterior predictive means.
```

We now compute the mean squared error for each of the four predictions:

```
. generate mse_y     = (y-yreg)^2
(100 missing values generated)
. generate mse_sw    = (y-ysw)^2
(100 missing values generated)
. generate mse_lasso = (y-ylasso)^2
(100 missing values generated)
. generate mse_bma   = (y-ybma)^2
(100 missing values generated)
. summarize mse*
```

Variable	Obs	Mean	Std. dev.	Min	Max
mse_y	100	1.315471	1.544705	.0006445	8.494073
mse_sw	100	1.295875	1.57022	.0000219	8.754056
mse_lasso	100	1.246921	1.507352	.0003377	7.452369
mse_bma	100	1.174436	1.375909	.0002316	5.69697

The BMA prediction has the lowest mean squared error. Of course, a proper comparison of the techniques requires a carefully designed simulation study.

◁

▷ Example 3: BMA with small sample size and many predictors, $n \leq p$

Let's now consider a case when the number of observations is too small relative to the number of predictors.

```
. use https://www.stata-press.com/data/r18/bmaintrosmall, clear
(Simulated data for BMA example, small sample)
. notes list y
y:
  1.  y = .5 + 1.2*x2 + 5*x10 + rnormal()
. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
y	10	.0976614	4.145433	-5.263075	6.823442
x1	10	-.2640087	1.147843	-2.680089	1.069156
x2	10	-.5486203	1.202882	-2.306713	1.269136
x3	10	.4727975	1.193019	-.9573489	3.020441
x4	10	-.0216079	.8695972	-1.82316	1.216567
x5	10	.2634739	.9095448	-1.28917	1.439632
x6	10	.091497	1.36508	-2.567606	1.809207
x7	10	.3522653	1.033754	-1.115946	1.885638
x8	10	-.1419826	.4697729	-.8331077	.6677282
x9	10	-.0343085	1.213427	-2.035336	1.647427
x10	10	-.0635723	.7551339	-1.023638	1.19934

In our toy example, we have only 10 observations, which is too small to make any reliable inferential conclusions. But we use it here for demonstration purposes to avoid dealing with too many variables. In practice, one can imagine datasets with, say, 100 observations and more than 100 potential predictors of which only a few are important in explaining the outcome, and we would like to investigate which ones. The analysis below can be easily adapted to datasets with more observations and variables.

Considering that the number of predictors in our dataset equals the sample size, we expect the traditional linear regression analysis and stepwise selection to fail. And they do.

```
. regress y x1-x10
note: x10 omitted because of collinearity.
```

Source	SS	df	MS	Number of obs	=	10
				F(9, 0)	=	.
Model	154.661546	9	17.1846162	Prob > F	=	.
Residual	0	0	.	R-squared	=	1.0000
				Adj R-squared	=	.
Total	154.661546	9	17.1846162	Root MSE	=	0

y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
x1	-4.475056	.	.	.	.	.
x2	2.618239	.	.	.	.	.
x3	-3.52965	.	.	.	.	.
x4	-3.814989	.	.	.	.	.
x5	-.1365321	.	.	.	.	.
x6	1.262926	.	.	.	.	.
x7	1.092976	.	.	.	.	.
x8	-2.792013	.	.	.	.	.
x9	-.7586842	.	.	.	.	.
x10	0	(omitted)				
_cons	1.051957	.	.	.	.	.

Because of insufficient sample size, `regress` arbitrarily omits one of the highest collinear predictors from the model because of collinearity. This happens to be one of the important predictors, `x10`. Also, as expected, `regress` fails to produce standard errors and p-values for the coefficients.

`stepwise` is not designed for $n \leq p$ and errors out.

We run `lasso linear` and compute predictions as before in example 2, except we use the same sample for fitting and prediction. Because of the small sample size, checking the out-of-sample (predictive) performance of the models is not feasible. Instead, we compare their in-sample performance, also known as model fit.

```
. lasso linear y x1-x10, rseed(18) nolog
Lasso linear model                          No. of obs        =         10
                                            No. of covariates =         10
Selection: Cross-validation                 No. of CV folds   =         10
```

ID	Description	lambda	No. of nonzero coef.	Out-of-sample R-squared	CV mean prediction error
1	first lambda	3.539901	0	-0.2057	18.64734
55	lambda before	.2871323	3	0.8906	1.692064
* 56	selected lambda	.2740817	3	0.8907	1.690632
57	lambda after	.2616242	3	0.8907	1.691032
60	last lambda	.2275474	3	0.8897	1.70528

```
* lambda selected by cross-validation.
. lassocoef, display(coef) nolegend
```

	active
x2	1.257819
x3	-.124988
x10	3.152851
_cons	0

```
. predict ylasso
(options xb penalized assumed; linear prediction with penalized coefficients)
```

The penalized coefficient of 3.15 for `x10` is not as close to the true value of 5. However, when the goal of the analysis is the optimal prediction, the actual coefficient estimates are of limited interest. And, in the context of `lasso`, it would not be appropriate to use these penalized coefficient estimates for inference anyway; see, for instance, [LASSO] **dsregress** instead.

We now fit BMA linear regression and compute predictions by using `bmaregress` and `bmapredict`, respectively.

```
. bmaregress y x1-x10
Enumerating models ...
Computing model probabilities ...
Bayesian model averaging                    No. of obs         =     10
Linear regression                           No. of predictors  =     10
Model enumeration                                     Groups   =     10
                                                      Always   =      0
Priors:                                     No. of models      =  1,024
  Models: Beta-binomial(1, 1)                   For CPMP >= .9 =     47
   Cons.: Noninformative                    Mean model size    =  2.967
   Coef.: Zellner's g
       g: Benchmark, g = 100                Shrinkage, g/(1+g) = 0.9901
  sigma2: Noninformative                    Mean sigma2        =  0.916
```

y	Mean	Std. dev.	Group	PIP
x10	4.785368	.7709731	10	.99683
x2	1.353152	.5137089	2	.94675
x3	-.1178808	.4227608	3	.18263
x1	.0877212	.5042626	1	.17811
x6	.0642453	.2037918	6	.15993
x8	-.1180912	.7904259	8	.12465
x9	.0469446	.2233004	9	.11361
x7	-.0404475	.2257238	7	.10327
x4	-.0364019	.4581988	4	.099553
x5	-.0046	.0954065	5	.062103
Always				
_cons	1.216777	.399357	0	1

```
Note: Coefficient posterior means and std. dev. estimated from 1,024 models.
Note: Default priors are used for models and parameter g.
. bmapredict ybma, mean
note: computing analytical posterior predictive means.
```

`bmaregress` still identifies the two important predictors, but the `PIP` estimates are now smaller—0.997 and 0.947 for `x10` and `x2`, respectively—compared with the values of 1 from example 2. This is expected given such a small sample size. In fact, a `PIP` as low as 0.5 would still qualify the predictor as important. The posterior mean estimates of the coefficients, 4.79 and 1.35, are reasonably close to their true values, 5 and 1.2, especially considering the small sample.

```
. generate mse_lasso = (y-ylasso)^2
. generate mse_bma   = (y-ybma)^2
. summarize mse*
```

Variable	Obs	Mean	Std. dev.	Min	Max
mse_lasso	10	.5736865	.7505244	.007729	2.541895
mse_bma	10	.3219035	.3677712	.0037272	1.186045

According to the smaller mean squared error, BMA produces predictions that are closer to the observed values than lasso in this example. And, unlike lasso, BMA can produce credible intervals for the predictions; see [BMA] **bmapredict**.

◁

It is difficult to generalize the conclusions based on these simple examples to other more complex situations, because we only looked at one dataset and one realization of the DGM. A proper simulation study is needed to make more general conclusions. But our limited findings appear to agree with some of the results reported in the literature.

Brief background and literature review

The initial development of the concept of model averaging was driven by the application problems, which have not been considered by mainstream statisticians. Barnard (1963) was one of the first to use a combination of models. His research was in quality-control methods with application to airline data. An early work by Bates and Granger (1969) introduced the idea of model combinations to problems of forecasting and influenced a string of follow-up articles, such as Newbold and Granger (1974) and Winkler and Makridakis (1983). During the 1970s, the development of model averaging took place mostly in economics research.

In statistical research, model averaging was also motivated by problems of prediction. Roberts (1965) viewed marginal distributions, either prior or posterior, as predictive distributions suitable to answer questions about model selection, interpretation, and validation. He suggested combining two models based on two different elicited priors. His idea was generalized by Leamer (1978), who was particularly interested in the uncertainty involved in model selection. Despite this early work, it took another two decades of the theoretical work for BMA to become a principled statistical method (Draper 1995; Kass and Wasserman 1995; and George 2014). Meanwhile, the developments in Bayesian computation, such as MCMC sampling methods, allowed researchers to effectively apply BMA in practice (Madigan and York 1995; Raftery 1996; Raftery, Madigan, and Volinsky 1995; and Hoeting et al. 1999). Madigan and Raftery (1994) showed the optimal predictive performance of BMA for high-dimensional contingency tables in comparison with model-selection methods. Clyde (1999) investigated prior specification and model search strategies in BMA.

The popularity of BMA in various scientific disciplines grew substantially. Fragoso, Bertoli, and Louzada (2018) provide a systematic review of published articles from 1996 to 2014. An in-depth survey of model-averaging application to problems of ecology is presented in Dormann et al. (2018). For application of BMA in political science, see Adams, Bishin, and Dow (2004) and Montgomery and Nyhan (2010).

The use of model-averaging methods in economic research remains strong. The application of BMA to problems of empirical microeconomics, with emphasis on big-data problems, is discussed in Koop (2017). A general overview of the use of model averaging in economics is given by Steel (2020). Among the questions in economic research, BMA has been traditionally applied to determining the growth factors driving economic processes (Brock and Durlauf 2001; Fernández, Ley, and Steel 2001b; Lenkoski, Eicher, and Raftery 2014; and Eicher and Newiak 2013). BMA is also a popular approach in policy and decision-making evaluation (Brock, Durlauf, and West 2003). The benefit of BMA as a tool for dealing with uncertainty in economic research is well documented in Marinacci (2015).

A survey of statistical methods accounting for model uncertainty demonstrates the advantage of BMA over other popular model-selection methodologies (Porwal and Raftery 2022). For comparison of BMA with other predictive methodologies, see Yao et al. (2018) and Piironen and Vehtari (2017).

References

Adams, J., B. G. Bishin, and J. K. Dow. 2004. Representation in congressional campaigns: Evidence for discounting/directional voting in U.S. Senate elections. *Journal of Politics* 66: 348–373. https://doi.org/10.1111/j.1468-2508.2004.00155.x.

Arin, K. P., and E. Braunfels. 2018. The resource curse revisited: A Bayesian model averaging approach. *Energy Economics* 70: 170–178. https://doi.org/10.1016/j.eneco.2017.12.033.

Atchadé, Y. F., and J. S. Rosenthal. 2005. On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* 11: 815–828. https://doi.org/10.3150/bj/1130077595.

Banner, K. M., and M. D. Higgs. 2017. Considerations for assessing model averaging of regression coefficients. *Ecological Applications* 27: 78–93. https://doi.org/10.1002/eap.1419.

Barnard, G. A. 1963. New methods of quality control. *Journal of the Royal Statistical Society, Series A* 126: 255–258. https://doi.org/10.2307/2982365.

Bates, J. M., and C. W. J. Granger. 1969. The combination of forecasts. *Operational Research* 20: 451–468. https://doi.org/10.2307/3008764.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24: 123–140. https://doi.org/10.1007/BF00058655.

Brock, W. A., and S. N. Durlauf. 2001. What have we learned from a decade of empirical research on growth? Growth empirics and reality. *World Bank Economic Review* 15: 229–272. https://doi.org/10.1093/wber/15.2.229.

Brock, W. A., S. N. Durlauf, and K. D. West. 2003. Policy evaluation in uncertain economic environments. *Brookings Papers on Economic Activity* 1: 235–322. https://doi.org/10.1353/eca.2003.0013.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: An integral part of inference. *Biometrics* 53: 603–618. https://doi.org/10.2307/2533961.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* 158: 419–466. https://doi.org/10.2307/2983440.

Clyde, M. A. 1999. Bayesian model averaging and model search strategies. In Vol. 6 of *Bayesian Statistics: Proceedings of the Sixth Valencia International Meeting*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 157–185. Oxford: Clarendon Press.

Darwen, P. J. 2019. Bayesian model averaging for river flow prediction. *Applied Intelligence* 49: 103–111. https://doi.org/10.1007/s10489-018-1232-0.

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

Dormann, C. F., J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M. Beale, S. Ciuti, J. Elith, K. Gerstner, J. Guelat, P. Keil, J. J. Lahoz-Monfort, L. J. Pollock, B. Reineking, D. R. Roberts, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, S. N. Wood, R. O. Wüest, and F. Hartig. 2018. Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs* 88: 485–504. https://doi.org/10.1002/ecm.1309.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* 57: 45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x.

——. 1999. Comment [on Hoeting et al. (1999)]. *Statistical Science* 14: 405–409.

Eicher, T. S., and M. Newiak. 2013. Intellectual property rights as development determinants. *Canadian Journal of Economics* 46: 4–22. https://doi.org/10.1111/caje.12000.

Eicher, T. S., C. Papageorgiou, and A. E. Raftery. 2011. Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics* 26: 30–55. https://doi.org/10.1002/jae.1112.

Fernández, C., E. Ley, and M. F. J. Steel. 2001a. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100: 381–427. https://doi.org/10.1016/S0304-4076(00)00076-2.

——. 2001b. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576. https://doi.org/10.1002/jae.623.

Fragoso, T. M., W. Bertoli, and F. Louzada. 2018. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review* 86: 1–28. https://doi.org/10.1111/insr.12243.

George, E. I. 2014. Bayesian model selection. In *Wiley StatsRef: Statistics Reference Online*, ed. N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels. New York: Wiley. https://doi.org/10.1002/9781118445112.stat00228.

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society, Series B* 14: 107–114. https://doi.org/10.1111/j.2517-6161.1952.tb00104.x.

Hansen, B. E. 2007. Least squares model averaging. *Econometrica* 75: 1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x.

Hansen, B. E., and J. S. Racine. 2012. Jackknife model averaging. *Journal of Econometrics* 167: 38–46. https://doi.org/10.1016/j.jeconom.2011.06.019.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14: 382–417. https://doi.org/10.1214/ss/1009212519.

Kaplan, D., and C. Lee. 2018. Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review* 42: 423–457. https://doi.org/10.1177/0193841X18761421.

Kass, R. E., and L. Wasserman. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90: 928–934. https://doi.org/10.1080/01621459.1995.10476592.

Koop, G. 2003. *Bayesian Econometrics*. Chichester, UK: Wiley.

——. 2017. Bayesian methods for empirical macroeconomics with big data. *Review of Economic Analysis* 9: 33–56. https://doi.org/10.15353/rea.v9i1.1434.

Leamer, E. E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Lenkoski, A., T. S. Eicher, and A. E. Raftery. 2014. Two-stage Bayesian model averaging in endogenous variable models. *Econometric Reviews* 33: 122–151. https://doi.org/10.1080/07474938.2013.807150.

Ley, E., and M. F. J. Steel. 2012. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* 171: 251–266. https://doi.org/10.1016/j.jeconom.2012.06.009.

Madigan, D., J. Gavrin, and A. E. Raftery. 1995. Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics—Theory and Methods* 24: 2271–2292. https://doi.org/10.1080/03610929508831616.

Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89: 1535–1546. https://doi.org/10.2307/2291017.

Madigan, D., and J. York. 1995. Bayesian graphical models for discrete data. *Journal of Statistical Review* 63: 215–232. https://doi.org/10.2307/1403615.

Marinacci, M. 2015. Model uncertainty. *Journal of the European Economic Association* 13: 1022–1100. https://doi.org/10.1111/jeea.12164.

Min, C., and A. Zellner. 1993. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* 56: 89–118. https://doi.org/10.1016/0304-4076(93)90102-B.

Montgomery, J. M., and B. Nyhan. 2010. Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis* 18: 245–270. https://doi.org/10.1093/pan/mpq001.

Moral-Benito, E. 2015. Model averaging in economics: An overview. *Journal of Economic Surveys* 29: 46–75. https://doi.org/10.1111/joes.12044.

Newbold, P., and C. W. J. Granger. 1974. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A* 137: 131–165. https://doi.org/10.2307/2344546.

Peisker, J. 2023. Context matters: The drivers of environmental concern in European regions. *Global Environmental Change* 79: 102636. https://doi.org/10.1016/j.gloenvcha.2023.102636.

Piironen, J., and A. Vehtari. 2017. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* 27: 711–735. https://doi.org/10.1007/s11222-016-9649-y.

Porwal, A., and A. E. Raftery. 2022. Comparing methods for statistical inference with model uncertainty. *PNAS* 119(16): e2120737119. https://doi.org/10.1073/pnas.2120737119.

Raftery, A. E. 1996. Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, ed. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, 163–187. Boca Raton, FL: Chapman and Hall.

Raftery, A. E., D. Madigan, and J. A. Hoeting. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92: 179–191. https://doi.org/10.1080/01621459.1997.10473615.

Raftery, A. E., D. Madigan, and C. T. Volinsky. 1995. Accounting for model uncertainty in survival analysis improves predictive performance. In Vol. 5 of *Bayesian Statistics: Proceedings of the Fifth Valencia International Meeting, June 5–9, 1994*, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 323–349. Oxford: Clarendon Press.

Raftery, A. E., and Y. Zheng. 2003. Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association* 98: 931–938. https://doi.org/10.1198/016214503000000891.

Roberts, H. V. 1965. Probabilistic prediction. *Journal of the American Statistical Association* 60: 50–62. https://doi.org/10.2307/2283136.

Steel, M. F. J. 2020. Model averaging and its use in economics. *American Economic Review* 58: 644–719. https://doi.org/10.1257/jel.20191385.

Tobias, J. L., and M. Li. 2004. Returns to schooling and Bayesian model averaging: A union of two literatures. *Journal of Economic Surveys* 18: 153–180. https://doi.org/10.1111/j.0950-0804.2004.00003.x.

Volinsky, C. T., D. Madigan, A. E. Raftery, and R. A. Kronmal. 1997. Bayesian model averaging in proportional hazards models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society, Series C* 46: 433–448. https://doi.org/10.1111/1467-9876.00082.

Wasserman, L. 2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology* 44: 92–107. https://doi.org/10.1006/jmps.1999.1278.

Winkler, R. L., and S. Makridakis. 1983. The combination for forecats. *Journal of the Royal Statistical Society, Series A* 146: 150–157. https://doi.org/10.2307/2982011.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5: 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman. 2018. Using stacking to average Bayesian predictive distributions (with dicscussion). *Bayesian Analysis* 13: 917–1007. https://doi.org/10.1214/17-BA10914.

Yin, G., and Y. Yuan. 2009. Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association* 104: 954–968. https://doi.org/10.1198/jasa.2009.ap08425.

Also see

[BMA] **BMA commands** — Introduction to commands for Bayesian model averaging

[BMA] **Glossary**

[BAYES] **Intro** — Introduction to Bayesian analysis

[BAYES] **Glossary**