Title

bmastats msize — Model-size summary after BMA regression

Description

`bmastats msize` provides a model-size summary after the `bmaregress` command. `bmastats msize` is useful to assess the overall complexity of the models in Bayesian model averaging (BMA) weighted by their prior and posterior model probabilities. By comparing the overall posterior model sizes with the prior model sizes, we can also assess the impact of the data on the BMA analysis.

Quick start

Display prior and posterior model-size summary

```
bmastats msize
```

Include a constant term in the model-size computations

```
bmastats msize, constant
```

Menu

Statistics > Bayesian model averaging > Model-size summary

Syntax

```
bmastats msize [, constant]
```

`collect` is allowed; see **[U] 11.1.10 Prefix commands**.

Option

Main

`constant` specifies that the constant term be included in model-size computations. By default, the constant term is not included.

Remarks and examples

Model size is the number of predictors included in a model. The constant term is typically not included in the model size. In BMA, there are multiple models, so we have a distribution of the model sizes. The model prior determines the prior model-size distribution. After you observe the data, the model prior is updated to form the posterior model-size distribution.

You can use the model-size distribution summaries to explore the complexity of the fitted BMA model. For instance, when the posterior median model size is small relative to the total number of regression terms, this means there are only a few strong predictors of the outcome. Conversely, when it is large, there are many weak potential predictors. And by comparing the posterior model size with the prior one, we can assess how the data affect our prior knowledge.

The prior model-size distribution can be computed analytically based on the known formula, with a caveat that in cases when the model space is not explored fully, this analytical distribution is conditional on the visited models.

The posterior model-size distribution is available analytically only with fixed g and is thus not provided with random g.

Whenever a posterior Markov chain Monte Carlo (MCMC) model sample is available, such as with random g or when bmaregress's sampling option is used, the posterior model-size distribution can be estimated from the sample by using sampling frequencies. See *Methods and formulas* for details.

bmastats msize reports prior and posterior model-size summaries, including mean and median model sizes. For prior model-size summary, it always provides the analytical estimates. For posterior model-size summary, it provides the analytical estimates with fixed g and the frequency estimates with random g. With fixed g and MC3 sampling, the command provides both types of estimates. By default, the constant is not included in the model-size computations, but you can specify the constant option to include it.

▷ Example 1: Model-size analysis after BMA regression using enumeration

Consider the performance dataset (Chatterjee and Hadi 2012, sec. 3.3) analyzed in example 1 of [BMA] **bmaregress**. The employees' satisfaction with their supervisors, rating, is modeled by six potential predictors. The total of 30 observations represent 30 different departments in the surveyed organization.

We fit a linear BMA regression using the bmaregress command.

```
. use https://www.stata-press.com/data/r18/performance
(Data on employee satisfaction with supervisor)

. bmaregress rating complaints-advance

Enumerating models ...
Computing model probabilities ...

Bayesian model averaging                           No. of obs         =     30
Linear regression                                  No. of predictors  =      6
Model enumeration                                              Groups =      6
                                                               Always =      0
Priors:                                            No. of models      =     64
  Models: Beta-binomial(1, 1)                          For CPMP >= .9 =     10
   Cons.: Noninformative                           Mean model size    =  1.699
   Coef.: Zellner's g
       g: Benchmark, g = 36                        Shrinkage, g/(1+g) = 0.9730
  sigma2: Noninformative                           Mean sigma2        = 52.302

------------------------------------------------------------------------------
      rating |      Mean   Std. dev.                          Group        PIP
-------------+----------------------------------------------------------------
  complaints |  .7052859   .1224289                               1     .99973
    learning |  .0603014   .1285281                               3     .25249
     advance | -.0167921    .073883                               6     .13148
  privileges | -.0074174   .0488635                               2     .10998
      raises |  .0070789   .0670475                               4     .10642
    critical |  .0009713   .0437848                               5    .098534
-------------+----------------------------------------------------------------
Always       |
       _cons |   14.8472   7.874219                               0          1
------------------------------------------------------------------------------
Note: Coefficient posterior means and std. dev. estimated from 64 models.
Note: Default priors are used for models and parameter g.
```

From the output, the model space is fully explored through enumeration. There is a total of $2^6 = 64$ models in the full model space.

Let's use bmastats msize to compute the model-size summary.

```
. bmastats msize

Model-size summary

Number of models = 64
Model size:
  Minimum = 0
  Maximum = 6

----------------------------------------
             |        Mean      Median
-------------+--------------------------
Prior        |
  Analytical |      3.0000           3
-------------+--------------------------
Posterior    |
  Analytical |      1.6986           1
----------------------------------------
Note: Frequency summaries not available.
```

The reported model size does not include the constant, so its range is between 0 and 6. The prior mean and median model sizes are both 3. The posterior mean and median model sizes are, respectively, 1.699 and 1. (The posterior mean model size is the same as the one reported by bmaregress.)

The default model prior distribution assumes a uniform prior on the model size. The prior mean model size, 3, is larger than the posterior one, 1.699. The posterior favors smaller models.

For the prior model-size summary, `bmastats msize` always reports analytical estimates. With model enumeration, it reports analytical estimates for the posterior model-size summary too. Frequency-based posterior estimates are not available here because there is no MCMC sample for the models—the models were enumerated.

◁

▷ Example 2: Model-size analysis after BMA regression using MC3 sampling

We fit the same BMA model as in example 1, but this time we use the MC3 sampling algorithm.

```
. bmaregress rating complaints-advance, sampling rseed(18)
Burn-in ...
Simulation ...
Computing model probabilities ...
Bayesian model averaging                          No. of obs         =     30
Linear regression                                 No. of predictors  =      6
MC3 sampling                                                  Groups =      6
                                                              Always =      0
                                                  No. of models      =     32
                                                      For CPMP >= .9 =     10
Priors:                                           Mean model size    =  1.699
  Models: Beta-binomial(1, 1)                     Burn-in            =  2,500
   Cons.: Noninformative                          MCMC sample size   = 10,000
   Coef.: Zellner's g                             Acceptance rate    = 0.2417
       g: Benchmark, g = 36                       Shrinkage, g/(1+g) = 0.9730
  sigma2: Noninformative                          Mean sigma2        = 52.292
Sampling correlation = 0.9990

------------------------------------------------------------------------------
      rating |      Mean   Std. dev.                               Group    PIP
-------------+----------------------------------------------------------------
  complaints |   .705479   .1218881                                    1      1
    learning |  .0601919   .1282869                                    3  .25234
     advance | -.0167514   .0737415                                    6  .13141
  privileges | -.0074265    .048844                                    2  .10996
      raises |  .0069949   .0666406                                    4  .10629
    critical |  .0009699   .0437742                                    5 .098526
-------------+----------------------------------------------------------------
Always       |
       _cons |  14.84478   7.871046                                    0      1
------------------------------------------------------------------------------
Note: Coefficient posterior means and std. dev. estimated from 32 models.
Note: Default priors are used for models and parameter g.
```

Instead of enumerating models, `bmaregress` generates a sample from the posterior model distribution that includes 32 different models.

```
. bmastats msize
Model-size summary

Number of models = 32
Model size:
  Minimum = 1
  Maximum = 6

─────────────┬──────────────────────────
             │        Mean      Median
─────────────┼──────────────────────────
Prior        │
  Analytical │      4.3333           5
─────────────┼──────────────────────────
Posterior    │
  Analytical │      1.6985           1
   Frequency │      1.7791           1
─────────────┴──────────────────────────
```

Here the minimum model size is 1, compared with 0 in example 1. This means that the null model, having a low posterior probability, was not visited by the MC3 sampler.

Although we used the same model prior as in example 1, the prior model-size estimates are different. This is because our explored model space now contains 32 models instead of all 64, and the prior model-size estimates are now conditional on the visited models.

With fixed g and when we fit a BMA model using MC3 sampling, in addition to analytical model-size estimates, the frequency estimates are also available. Provided that the model-space sampling converges, the analytical and frequency estimates should be close. In our example, the analytical and frequency model-size estimates, 1.7 and 1.8, are close.

We can also explore the effect of the g parameter on the complexity of our BMA model. Let us, for example, fix g to 1, which is much lower than the default value of 36 used above.

```
. bmaregress rating complaints-advance, gprior(fixed 1) sampling rseed(18)
Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                          No. of obs         =      30
Linear regression                                 No. of predictors  =       6
MC3 sampling                                                  Groups =       6
                                                              Always =       0
                                                  No. of models      =      63
                                                     For CPMP >= .9 =       29
                                                  Mean model size    =   3.731
Priors:                                           Burn-in            =   2,500
  Models: Beta-binomial(1, 1)                     MCMC sample size   =  10,000
   Cons.: Noninformative                          Acceptance rate    =  0.5678
   Coef.: Zellner's g                             Shrinkage, g/(1+g) =  0.5000
       g: g = 1                                   Mean sigma2        = 103.952
  sigma2: Noninformative

Sampling correlation = 0.9957

------------------------------------------------------------------------------
      rating |      Mean   Std. dev.                          Group        PIP
-------------+----------------------------------------------------------------
  complaints |  .3188165    .158556                               1     .94416
    learning |  .0944454   .1508832                               3     .61347
     advance | -.0520093    .137935                               6     .55788
      raises |   .030129   .1654381                               4      .5473
  privileges | -.0163312   .1027766                               2     .53808
    critical |  .0062495   .1066877                               5     .52978
-------------+----------------------------------------------------------------
Always       |
       _cons |  38.76265   10.36275                               0          1
------------------------------------------------------------------------------
Note: Coefficient posterior means and std. dev. estimated from 63 models.
Note: Default prior is used for models.

. bmastats msize

Model-size summary

Number of models = 63
Model size:
  Minimum = 0
  Maximum = 6

----------------------------------------
             |           Mean     Median
-------------+--------------------------
Prior        |
  Analytical |         3.0488          3
-------------+--------------------------
Posterior    |
  Analytical |         3.7307          4
   Frequency |         3.6529          4
----------------------------------------
```

The posterior mean model size has increased to 3.73, and the posterior median model size has increased to 4. With $g = 1$, BMA appears to favor larger models.

◁

▷ Example 3: Model-size analysis after BMA regression with random g

Both example 1 and example 2 used a fixed g. Let's explore the case of a random g. (An in-depth coverage of the effects of the g-prior on model complexity can be found in, for example, Ley and Steel [2012].)

To demonstrate, we will use a `robust` prior for g.

```
. bmaregress rating complaints-advance, gprior(robust) rseed(18)
Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                            No. of obs        =     30
Linear regression                                   No. of predictors =      6
MC3 and adaptive MH sampling                                   Groups =      6
                                                               Always =      0
                                                    No. of models     =     34
                                                       For CPMP >= .9 =     12
Priors:                                             Mean model size   =  1.734
  Models: Beta-binomial(1, 1)                       Burn-in           =  2,500
   Cons.: Noninformative                            MCMC sample size  = 10,000
   Coef.: Zellner's g                               Acceptance rate   = 0.4232
       g: Robust
  sigma2: Noninformative                            Mean sigma2       = 53.095

Sampling correlation = 0.9994

------------------------------------------------------------------------------
      rating |       Mean  Std. dev.                            Group      PIP
-------------+----------------------------------------------------------------
  complaints |   .7000463  .1273543                                1    .9998
    learning |   .0594904  .1286095                                3      .25
     advance |  -.0192712  .0797935                                6    .1503
      raises |   .0079416  .0727859                                4    .1201
  privileges |  -.0072591  .0487009                                2    .1069
    critical |   .0014397  .0466476                                5    .1067
-------------+----------------------------------------------------------------
Always       |
       _cons |   15.24911  7.988166                                0        1
------------------------------------------------------------------------------
Note: Coefficient posterior means and std. dev. estimated from 34 models.
Note: Default prior is used for models.

------------------------------------------------------------------------------
             |                                              Equal-tailed
             |       Mean  Std. dev.      MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
           g |    152.668   1968.132   43.5265   33.81024   8.205076   610.6026
   Shrinkage |   .9656427   .0276071   .001234   .9712728   .8913639   .9983649
------------------------------------------------------------------------------
```

bmaregress now uses MC3 sampling for the models and an adaptive Metropolis–Hastings sampling for g.

```
. bmastats msize

Model-size summary

Number of models = 34
Model size:
  Minimum = 1
  Maximum = 6

─────────────┬─────────────────────
             │       Mean    Median
─────────────┼─────────────────────
Prior        │
  Analytical │     4.1786         4
─────────────┼─────────────────────
Posterior    │
   Frequency │     1.7338         1
─────────────┴─────────────────────
Note: Analytical summaries not available.
```

Analytical posterior estimates are not available with random g. The frequency posterior estimates are similar to those in example 2 for fixed $g = 36$.

◁

Stored results

bmastats msize stores the following in r():

Scalars	
r(k_models)	number of models
r(msize_mean_prior)	prior mean model size
r(msize_mean_a)	analytical posterior mean model size (not available with random g)
r(msize_mean_f)	frequency posterior mean model size (not available with model enumeration)
r(constant)	1 if constant is specified; 0 otherwise
Matrices	
r(modelsize)	model-size summary

Methods and formulas

Consider a BMA regression model for an outcome vector $\mathbf{y}$ with p predictors. Let $\mathcal{M}_F = \{M_1, M_2, \dots, M_{2^p}\}$ denote the full space of models formed by considering all 2^p possible subsets of p variables, and let $J_F = \{1, 2, \dots, 2^p\}$ denote the full set of the corresponding indices. Let $|M|$ denote the model size of a regression model M from $\mathcal{M}_F$; that is, $|M|$ equals the number of predictors included in model M.

The prior distribution of $|M|$ is a discrete distribution on the set $\{0, 1, \dots, p\}$ such that

$$\Pr(|M| = s) = \sum_{j=1}^{2^p} I(|M_j| = s) P(M_j)$$

where $I(\cdot)$ is an indicator function, M_j's are all possible enumerated models from $\mathcal{M}_F$, and $P(M_j)$'s are the prior model probabilities.

Similarly, the posterior distribution of $|M|$ is defined as

$$\Pr(|M| = s|\mathbf{y}) = \sum_{j=1}^{2^p} I(|M_j| = s) P_a(M_j|\mathbf{y})$$

where $P_a(M_j|\mathbf{y})$'s are the analytical posterior model probabilities defined by (7) in *Posterior model probability* in *Methods and formulas* of [BMA] **bmaregress**.

When the model space is fully explored through enumeration, the analytical prior mean model size is

$$E(|M|) = \sum_{j=1}^{2^p} |M_j| P(M_j)$$

and the analytical posterior mean model size is

$$E(|M|\,|\,\mathbf{y}) = \sum_{j=1}^{2^p} |M_j| P_a(M_j|\mathbf{y})$$

When model sampling is used instead of model enumeration, the analytical prior mean model size is estimated conditionally on the subspace of the visited models indexed by $J \subset J_F$:

$$\widehat{E}(|M|) = \frac{\sum_{j\in J} |M_j| P(M_j)}{\sum_{j\in J} P(M_j)}$$

With fixed g and when model sampling is used, such as when `bmaregress`'s `sampling` option is specified, the analytical posterior mean model size is estimated as

$$\widehat{E}_a(|M|\,|\,\mathbf{y}) = \frac{\sum_{j\in J} |M_j| P_a(M_j|\mathbf{y})}{\sum_{j\in J} P_a(M_j|\mathbf{y})}$$

With random g, analytical formulas for posterior model probabilities and posterior model-size probabilities are not available.

When a posterior MCMC sample of models, $\{m_t\}_{t=1}^T$, is available, such as with random g or when `bmaregress`'s `sampling` option is used, the frequency estimate of the posterior mean model size is computed as follows:

$$\widehat{E}_f(|M|\,|\,\mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} |m_t|$$

References

Chatterjee, S., and A. S. Hadi. 2012. *Regression Analysis by Example*. 5th ed. New York: Wiley.

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.

Also see

[BMA] **bmagraph msize** — Model-size distribution plots after BMA regression

[BMA] **bmastats** — Summary for models and predictors after BMA regression

[BMA] **bmaregress** — Bayesian model averaging for linear regression

[BMA] **BMA postestimation** — Postestimation tools for Bayesian model averaging

[BMA] **Glossary**

For suggested citations, see the FAQ on citing Stata documentation.