Luis Huesca (CIAD) / Enrique Labrada del Razo (UABC)

Data management in household income and expenditure surveys: Working with extended families using Stata.

Introduction

What is a model averaging (MA)?

- The method of MA has become an important tool to deal with model uncertanty (Steel, 2020).

- Instead of relying on just one model, MA averages results over multiple plausible models based on the observed data.

- In statistical analysis, to select the right model is a primary and crucial step. So, in predictive inference, single-model approaches do not use all the information that is available which make the inference unstable (StataCorp, 2023).

- The difference in the MA approach is based on the fact that instead of selecting just one model, it considers a list of candidate models. 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.

- So the larger the candidate model space is, the greater the possibility of model may change every time the new data become available (StataCorp, 2023).

Background and brief review

- In the need to modeling selection in one time –currently still applied by novel students and obsolete techniques among professionals are used as well- to fulfil and get the best regression models.
- The Bayesian model choice (BMC) & Bayesian Model Averaging (BMA) with application to variable selection is shown by Bayarri, Berger, Forte, and García-Donato (2012).
- Recent studies were carried out inspired by using previous approaches by Foster and George (1994) ; Kass and Wasserman (1995); Madigan and York (1995) research.
- Last year Porwal and Raftery (2022) has created a widely spectrum for comparing BMC methods for statistical inference with model uncertainty.

What is Bayesian model averaging (BMA)?

It is an application of Bayesian inference to the problems of model selection, combined estimation and prediction that produces a straightforward model choice criteria and less risky predictions (Fragoso & Louzada, 2015).

Bayes theorem

- It is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances.
- Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.

Bayes theorem

$$P(M|D) = \frac{P(D|M)\,P(M)}{\sum_{M^*} P(D|M^*)\,P(M^*)} \qquad (1)$$

- Model M: It *is a random variable with prior P(M) distributed over some model space.*
- D: Observed data.
- P(M): Likelihood of M.
- $P(D|M)$: It is the probability of D with respect to M, known as the marginal likelihood of model M.
- $P(M|D)$: It is known as the posterior model probability and is a key quantity in BMA inference and prediction.

Bayesian Model Averaging

- Basically, is known as model choice, parameter estimation, and prediction.
- BMA 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.

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”. Basically, is known as model choice, parameter estimation, and prediction.
- 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.

  - The importance of a predictor is based on the estimated **Posterior Inclusion Probability** (PIP), the probability that this predictor is included in a model estimated over the considered model space.
- 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.

Usage of BMA

- **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." (StataCorp, 2023).
- **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." (StataCorp, 2023).

BMA Comands & applications

○ **Setup**

- Splitsample: Split data into random samples for training, validation, and prediction

- vl: Manage large variable lists conveniently

○ **Estimation**

- bmaregress: BMA linear regression

- bmacoefsample: Posterior samples of regression coefficients

○ **Graphical commands**

- bmagraph: Graphical summaries

- bmagraph pmp: Model probability plots

- bmagraph varmap: Variable- inclusion map

- bmagraph msize: Model- size distribution plots

- bmagraph coefdensity: Coefficient density plots

- ○ **Postestimation statistics**

- bmastats: Posterior summaries

- bmastats msize: Model-size summary

- bmastats models: Posterior model and variable- inclusion summaries

- bmastats pip: Posterior inclusion probabilities foir prediction

- bmastats jointness: Jointness measures for predictors

- bmastats lps: Log predictive- score

- ○ **Predictions**

- bmapredict: BMA predictions

Sintax

BMA linear regression with in-out predictors

- bmaregress depvar [inoutvars] [if] [in] [weight] [, mprior(mspec) gprior(gspec) options]

BMA linear regression with always-included predictors

- bmaregress depvar (alwaysvars, always) [inoutvars] [if] [in] [weight] [,mprior(mspec) gprior(gspec) options]

BMA linear regression with groups of predictors

- bmaregress depvar [(alwaysvars, always)] [inoutspec] [if] [in] [weight] [, mprior(mspec) gprior(gspec) options]

Where `inoutvars` and `alwaysvars` are varlist.

Bmaregress-BMA empirical application by using MEXMOD database

Our goal:

1. Finding the appropriate model that contributes to determine the effect from non-contributory pensions on the labor supply of extended families.

- Proxy for labor supply are hours of work
- Run the next regression by using BMA:

```
. bmaregress lhw - lsic, sampling mprior(uniform) gprior(hyperg 3) rseed(18) dots saving(bmasim, replace)
. bmacoefsample
. bmagraph coefdensity {logylab dgn yem les_r dag} ,combine(rows(2))
. bmagraph coefdensity {i.rel} , combine(rows(2)) name(g1, replace)
```

2. Run model specification with OLS

3. Run model specification with Probit

4. Comparisson between the models.

Summary statistics

Table 1. Summary statistics for the BMA choice model, Mexico 2020

	Summary
N	83,039
Weekly working hours	41.693 (19.448)
RECODE of parentesco_n (Parentesco)	
Padre/Madre	14,546 (17.9%)
Conyuge	8,970 (11.0%)
Hijo	27,055 (33.3%)
Abuelo	2,674 (3.3%)
nuera/yerno	5,636 (6.9%)
nieto(a)	22,346 (27.5%)
dag	31.025 (22.256)
logb	7.003 (0.434)
logylab	8.077 (1.313)
logtax	4.378 (2.320)
RECODE of les	
Preescolar	3,524 (4.2%)
Trabajador del campo	2,530 (3.0%)
Empresario o autoempleado	5,601 (6.7%)
Empleado	27,936 (33.6%)
Pensionado	2,181 (2.6%)
Desempleado	1,981 (2.4%)
Estudiante	14,995 (18.1%)
Inactivo	15,599 (18.8%)
Enfermo o Incapacitado	1,647 (2.0%)
Otro	7,045 (8.5%)
boa_d	0.054 (0.226)
dgn	0.451 (0.498)
dru	0.393 (0.488)
deh	1.726 (1.541)
yem	1,820.842 (4,185.416)
yse	403.964 (3,453.368)
boa	67.399 (307.427)
lsic	4.606 (0.976)

Source: Own estimates using dtable command in Stata 18.

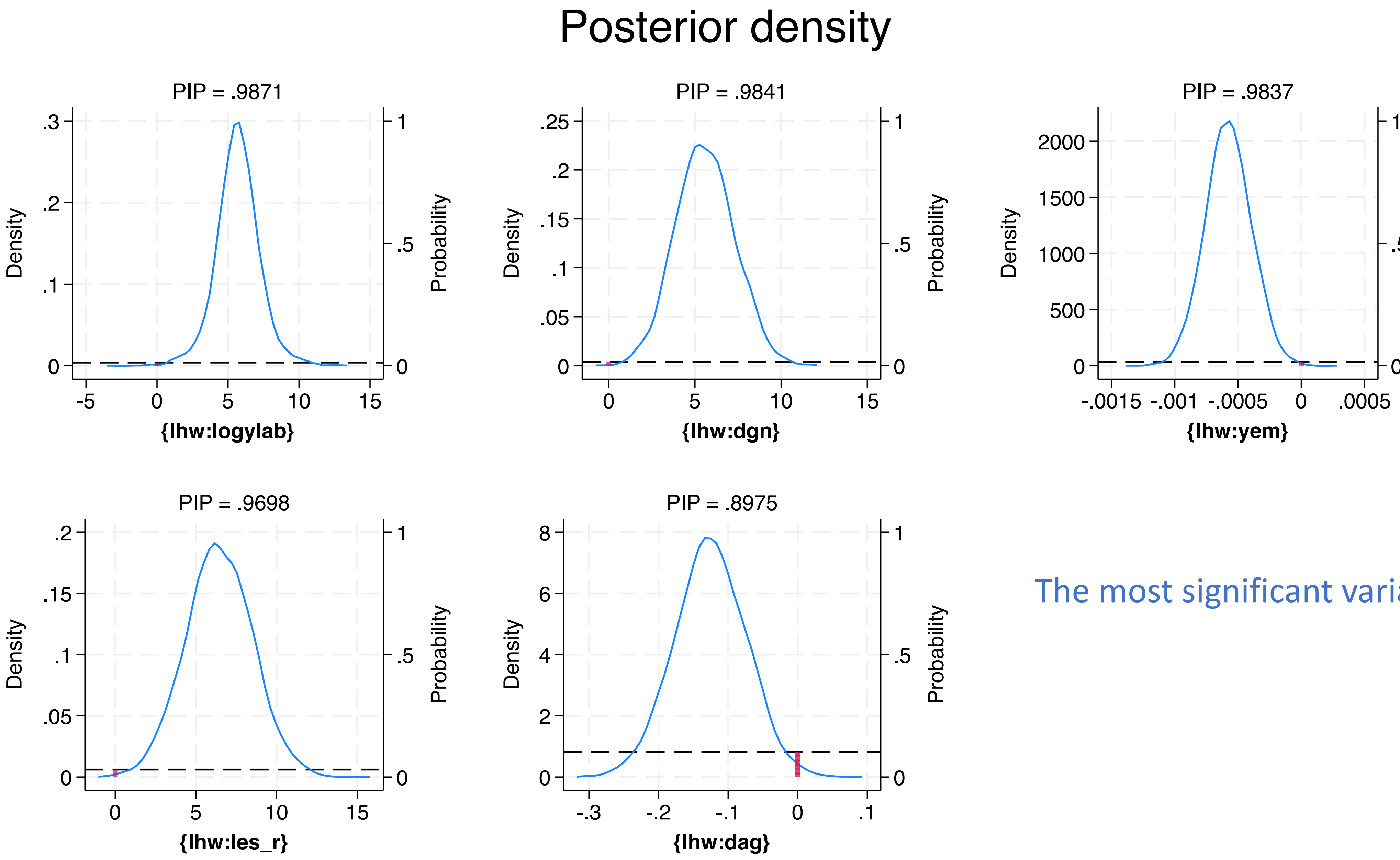
BMA Regression

```
Computing model probabilities ...

Bayesian model averaging                  No. of obs         =        867
Linear regression                         No. of predictors  =         18
MC3 and adaptive MH sampling                          Groups =         18
                                                      Always =          0
                                          No. of models      =      2,756
                                             For CPMP >= .9  =      1,775
Priors:                                   Mean model size    =      9.280
  Models: Uniform                         Burn-in            =      2,500
   Cons.: Noninformative                  MCMC sample size   =     10,000
   Coef.: Zellner's g                     Acceptance rate    =     0.7063
       g: Hyper-g(3)
  sigma2: Noninformative                  Mean sigma2        =    290.567

Sampling correlation = 0.6409
```

lhw	Mean	Std. dev.	Group	PIP
logylab	5.605452	1.640189	8	.9871
dgn	5.50595	1.849746	12	.9841
yem	-.0005669	.0001972	15	.9837
les_r	6.240007	2.343197	10	.9698
dag	-.111921	.0614636	6	.8975
rel				
Abuelo	-4.396838	4.41132	3	.6627
logb	-.9972583	1.650244	7	.4598
rel				
Hijo	.4837455	1.155486	2	.3757
boa_d	.8416644	2.438669	11	.3618
dru	-.2975518	.8728791	13	.318
rel				
nieto(a)	.5239862	2.402818	5	.3089
lsic	.0772457	.6111674	18	.3087
rel				
Conyuge	-.3818068	1.263674	1	.3025
boa	.0002677	.0016854	17	.292
logtax	.0092851	.7412757	9	.2822
yse	-2.23e-06	.0000326	16	.2699
deh	-.0080147	.2290189	14	.258
rel				
nuera/yerno	-.2317291	1.432597	4	.2577
Always				
_cons	-13.67939	20.49186	0	1

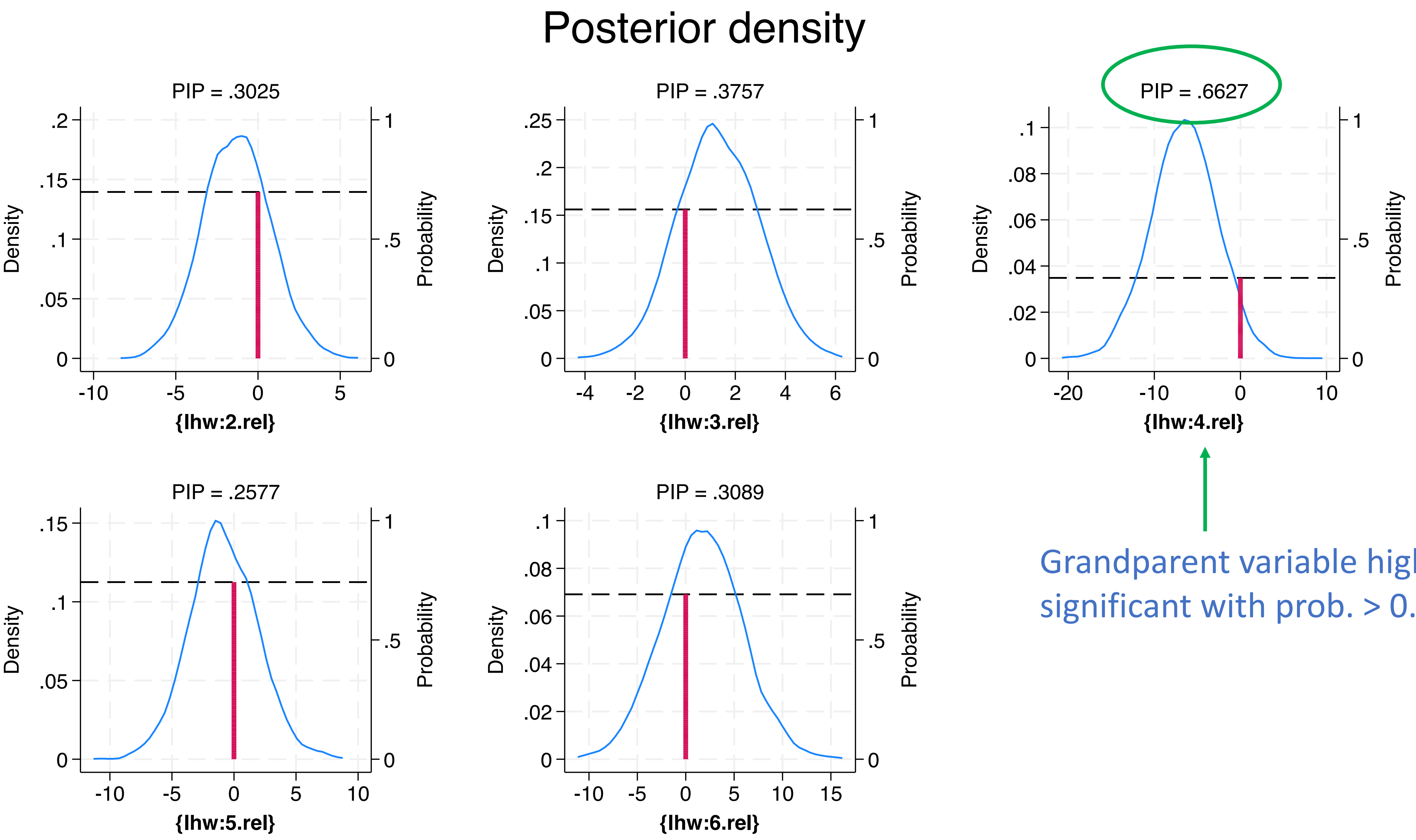
Note: Coefficient posterior means and std. dev. estimated from 2,756 models.

BMA graphical analysis



The most significant variables.

BMA graphical analysis



Grandparent variable highly significant with prob. > 0.5

Full model using OLS

horas	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
rel						
Conyuge	-1.171783	2.299102	-0.51	0.610	-5.68438	3.340815
Hijo	.880557	1.851388	0.48	0.634	-2.753284	4.514398
Abuelo	-7.694282	4.125391	-1.87	0.063	-15.79146	.4028925
nuera/yerno	-.9356803	3.094397	-0.30	0.762	-7.009256	5.137895
nieto(a)	1.89854	4.55498	0.42	0.677	-7.041818	10.8389
dag	-.1404735	.0643077	-2.18	0.029	-.2666944	-.0142527
logb	-2.48736	1.976902	-1.26	0.209	-6.367554	1.392834
logylab	6.746648	2.399165	2.81	0.005	2.037649	11.45565
logtax	-.2131222	1.398704	-0.15	0.879	-2.95845	2.532206
les_r	7.461944	2.469954	3.02	0.003	2.614004	12.30988
boa_d	3.660248	6.178854	0.59	0.554	-8.467392	15.78789
dgn	7.176366	1.869762	3.84	0.000	3.506461	10.84627
dru	-1.325468	1.472285	-0.90	0.368	-4.215218	1.564281
deh	-.0757605	.4924994	-0.15	0.878	-1.042421	.8909002
yem	-.0006973	.0002003	-3.48	0.001	-.0010905	-.0003041
yse	-8.91e-06	.0000683	-0.13	0.896	-.000143	.0001252
boa	-.0005762	.0047136	-0.12	0.903	-.0098279	.0086754
lsic	.5451472	1.136144	0.48	0.631	-1.684837	2.775131
_cons	-17.23632	22.44101	-0.77	0.443	-61.28276	26.81012

High prob. error in estimated coefficients.

Linear and probit regressions

Both regressions show highly significant estimated coefficients:

	Model 1: OLS	Model 2: Probit
main		
logylab	5.524***	-0.167***
dgn	6.473***	0.0787***
yem	-0.000326***	0.00000623***
2.les_r	0	0
3.les_r	2.352***	0.345***
4.les_r	6.127***	-0.0614
8.les_r	-4.868	0.462*
dag	-0.0596***	-0.00114
1.rel	0	0
2.rel	-1.929***	0.0816**
3.rel	-1.151***	-0.0524*
4.rel	0.761	0.00769
5.rel	0.588	0.0614*
6.rel	-3.245***	-0.0747
_cons	-5.895***	1.053***
N	32449	32449

* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Source: Own estimation by using esttab command written by Ben Jann.

Conclusions

- BMA new command for linear regression accounts for the uncertainty of which predictors should be included in the regression model.
- It can be used for inference, prediction, or model selection.
- Inference can be made about models based on posterior model probabilities (PMPs), importance of predictors based on posterior inclusion probabilities (PIPs), and regression coefficients based on their posterior distributions.
- bmaregress allows you to include predictors as groups and provides several ways of dealing with interaction terms. It supports a variety of priors for models and regression coefficients.
- Last but not least, good to recall that high PIP values, commonly above 0.5, are considered as important predictors.

References

Steel, M. F. J. (2020). Model averaging and its use in economics. American Economic Review 58: 644–719.

StataCorp. (2023). Stata: Release 18. Statistical Software. College Station, TX: StataCorp LLC: 1- 50.

Bayarri, M. J., J. O. Berger, A. Forte, and G. García-Donato (2012). Criteria for Bayesian model choice with application to variable selection. Annals of Statistics 40: 1550–1577.

Foster, D. P., and E. I. George (1994). The risk inflation criterion for multiple regression. Annals of Statistics 22:1947–1975.

Jann, Ben & Carlos, Alvarado (2014). Making regression tables from stored estimates. The Stata Journal, 14, Number 2, p. 451.

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.*

Madigan, D., and J. York (1995)*. Bayesian graphical models for discrete data. Journal of Statistical Review 63: 215–232.*

Porwal, A., and A. E. Raftery. (2022a)*. Effect of model space priors on statistical inference with model uncertainty. New England Journal of Statistics in Data Science 1–10.*

Fragoso, T. M., W. Bertoli, and F. Louzada. (2018)*. Bayesian model averaging: A systematic review and conceptual classification. International Statistical Review 86: 1–28.*