Bayesian Econometrics

Model Averaging: Theory and Applications

Andrés Ramírez Hassan

Universidad EAFIT

May 11, 2017

Outline

Outline

Introduction

Econometrics is mainly concerned with the construction of models to study economic phenomena. In frequentist methods, misspecification is a problem that is not addressed oftenly due to the difficulties it entails. The Bayesian paradigm, allows to treat the model specification as a random variable as well and therefore, a likelihood and prior can be formulated in order to deal with the issue of misspecification. This is done with a technique known as *Bayesian Model Averaging (BMA)*, that makes use of *Markov Chain Monte Carlo Model Composition (MC^3)* methods. This presentation will follow Koop (2003).

Motivation

Say one acquires data y and that there are R different models; one can think about attacking the specific problem. Then each model M_r with $r = 1, \ldots, R$, depends on a specific set of parameters, θ_r, and is characterized by a likelihood $p(y|\theta_r, M_r)$, a prior $p(\theta_r|M_r)$ and a posterior $p(\theta_r|y, M_r)$.

To learn about the parameters in each model we use Bayes' theorem

$$p(\theta_r|y, M_r) = \frac{p(y|\theta_r, M_r)p(\theta_r|M_r)}{p(y|M_r)} \tag{1}$$

Motivation

Since the model can also be thought of as being uncertain, we find *posterior model probabilities*

$$p(M_r|y) = \frac{p(y|M_r)p(M_r)}{p(y)} \tag{2}$$

Here $p(M_r)$ is known as the *prior model probability*. In order to find $p(y|M_r)$, the *marginal likelihood*, we can use Eq. 1

$$p(y|M_r) = \int p(y|\theta_r, M_r)p(\theta_r|M_r)d\theta_r \tag{3}$$

Motivation

When formulating a specific problem, one usually has some coefficients ϕ in mind that are common to all models. All of the information about ϕ is contained in $p(\phi|y)$ which can be found through simple rules of probability as

$$p(\phi|y) = \sum_{r=1}^{R} p(\phi|M_r, y)p(M_r|y) \tag{4}$$

Or, if $g(\phi)$ is a function of interest (such as the mean), then

$$E[g(\phi)|y] = \sum_{r=1}^{R} E[g(\phi)|M_r, y]p(M_r|y) \tag{5}$$

Motivation

In any case, one needs to compute $E[g(\phi)|M_r, y]$ and the posterior model probabilities $p(M_r|y)$ for every model. However, we know that both of these calculations can be strenuous and difficult to handle even if R remained small.

The literature has focused on cases where the quantities can be found analytically and techniques such as MC^3 have been developed to deal with a large number of models.

Outline

Overview

Different models can be derived in many ways. For example:

- A model with normally distributed errors against a t-student distributed errors.
- A logit against a probit model.
- A linear regression with different explanatory variables

In this presentation we focus on a normal linear regression model that is defined by its explanatory variables. Therefore, if there are K different possible variables, the number of models is $R = 2^K$. We start by discussing the likelihood, priors, posteriors and marginal likelihoods for this problem.

Likelihood

Formally, we say we have data on $i = 1, \ldots, N$ individuals. The dependent variable y is a $N \times 1$ vector and we have K possible independent variables, collected in the $N \times K$ matrix, X. Therefore, our models are M_r with $r = 1, \ldots, 2^K$. We assume there is an intercept α in each model and thus

$$y = \alpha \iota_N + X_r \beta_r + \epsilon \tag{6}$$

Where ι_N is a N-vector of ones, X_r is a $N \times k_r$ matrix containing some (or all) columns of X. ϵ is a N-vector assumed to be normally distributed with mean 0_N and covariance matrix $h^{-1} I_N$.

Likelihood

We can transform the distributional assumptions on ϵ to a distribution for y using the change of variables theorem. As such, y follows a multivariate normal distribution and the likelihood function is

$$p(y|\alpha, \beta_r, h; X_r, M_r) = \frac{h^{N/2}}{(2\pi)^{N/2}} \times \exp\left\{-\frac{h}{2}(y - \alpha\iota_N - X_r\beta_r)'(y - \alpha\iota_N - X_r\beta_r)\right\} \tag{7}$$

Prior

The prior distribution is extremely important for BMA and should be treated with care. For example, even if one had a lot of prior information, formulating 2^K different prior distributions would be next to impossible for a researcher. Therefore, we assume that the parameters that are common for each model follow the same prior distribution and set these to

$$p(h) \propto \frac{1}{h} \tag{8}$$

$$p(\alpha) \propto 1 \tag{9}$$

These are noninformative priors for both h and α.

Prior

For the β_r must be used informative priors. Non–informative priors are dangerous. These favor parsimony, no matter the data, or when the number of regressors are the same, posterior odds depend upon units of measure. So, we assume the usual conjugate priors, i.e. a normal distribution conditional on h such as

$$\beta_r|h \sim \mathcal{N}(\underline{\beta}_r, h^{-1}\underline{V}_r) \tag{10}$$

We center the prior mean around 0 and assume what is called a g-prior for $\underline{V}_r$. That is,

$$\underline{\beta}_r = 0_{k_r} \tag{11}$$

$$\underline{V}_r = (g_r X_r' X_r)^{-1} \tag{12}$$

Prior

This g-prior depends on data information (similar to empirical Bayes methods). It also helps reducing the selection of hyperparameters from $N(N+1)/2$ variance and covariances to a single g_r. This is also a benchmark prior used extensively in BMA.

Thus, the prior for each β_r is

$$\beta_r|h \sim \mathcal{N}(0_{k_r}, h^{-1}(g_r X_r' X_r)^{-1}) \tag{13}$$

Posterior

The posterior distributions will simply be a generalization of

$$\beta_r|h, y, M_r \sim \mathcal{N}(\overline{\beta}_r, h^{-1}\overline{V}) \tag{14}$$

$$h|y, M_r \sim \mathcal{G}(\overline{v}, \overline{s}^{-2}) \tag{15}$$

For our particular case, we find that

$$\overline{V} = [(1 + g_r)X_r'X_r]^{-1} \tag{16}$$

$$\overline{\beta}_r = \overline{V}X_r'y \tag{17}$$

$$\overline{s}^2 = \frac{1}{1 + g_r}y'P_{X_r}y + \frac{g_r}{1 + g_r}(y - \overline{y}\iota_N)'(y - \overline{y}\iota_N)/\overline{v} \tag{18}$$

$$P_{X_r} = I_N - X_r(X_r'X_r)^{-1}X_r' \tag{19}$$

$$\overline{v} = N \tag{20}$$

Marginal likelihood

Remember that the marginal likelihood is

$$p(y|M_r) = \int p(y|\theta_r, M_r)p(\theta_r|M_r)d\theta_r \tag{21}$$

For our particular case, this turns out to be

$$p(y|M_r) \propto \left(\frac{g_r}{1+g_r}\right)^{\frac{k_r}{2}} \tag{22}$$

$$\times \left[\frac{1}{1+g_r}y'P_{X_r}y + \frac{g_r}{1+g_r}(y-\overline{y}\iota_N)'(y-\overline{y}\iota_N)\right]^{-\frac{N-1}{2}} \tag{23}$$

Posterior model probabilities

We can compute these probabilities for each model analytically as

$$p(M_r|y) = cp(y|M_r)p(M_r) \tag{24}$$

Where c is just a constant that gets canceled in every comparison. Setting the prior model probability equal for every model to $1/R$, we just compute the Bayes factor of the likelihoods, normalizing

$$p(M_r|y) = \frac{p(y|M_r)}{\sum_{j=1}^{R} p(y|M_j)} \tag{25}$$

Outline

Motivation

When the number of models is high, it is extremely hard to compute the probabilities for all models and MC^3 methods were invented to deal with such cases. In particular, the algorithm by Madigan et. al (1995) generalizes the Metropolis-Hastings framework for model selection. First, you have a model in the previous iteration, say $M^{(g-1)}$. Then, a new proposed model M^* is drawn from the model space comprised by all the models that add, subtract or leave the same variables as $M^{(g-1)}$, with equal probability. The acceptance probability takes the form of

$$\alpha(M^{(g-1)}, M^*) = min\left\{\frac{p(y|M^*)p(M^*)}{p(y|M^{(g-1)})p(M^{(g-1)})}, 1\right\} \tag{26}$$

Posterior model probabilities

If we consider equal priors for both models, i.e. $p(M^*) = p(M^{(g-1)})$, these cancel out and we just need to compute the Bayes factor associated to them

$$BF = \frac{p(y|M^*)}{p(y|M^{(g-1)})} \tag{27}$$

This will be the fundamental calculation. By simulation, we can approximate the functions of parameters with the formula in Eq. (5).

Outline

Algorithm

1. Choose a starting model $M^{(0)}$.
2. At the gth iteration, draw a model from a similar model space as a candidate, say M^*
3. Compute the marginal likelihood $p(y|M_r)$ for $M^{(g)}$ and M^*
4. Calculate

$$\alpha(M^{(g-1)}, M^*) = min\left\{\frac{p(y|M^*)}{p(y|M^{(g-1)})}, 1\right\} \tag{28}$$

5. Generate U from $U(0,1)$
6. If $U \leq \alpha(M^{(g-1)}, M^*)$ then $M^{(g)} = M^*$, if not $M^{(g)} = M^{(g-1)}$ and go back to 1.

Simulation setting

There are five possible variables to choose from and the real model includes only the first three. There are $2^5 = 32$ possible models.

$$
\begin{aligned}
X &= [x_1; x_2; x_3; x_5; x_5] \\
x_i &\sim \mathcal{N}(0, (i+1)^2), i = 1, \ldots, 5 \\
\beta' &= [2, 3, 4] \\
\mu &\sim \mathcal{N}(0, 1) \\
y &= X^* \beta + \mu \\
X^* &= [x_1; x_2; x_3] \\
g_r &= \begin{cases} \frac{1}{N} & \text{if } N > K^2 \\ \frac{1}{K^2} & \text{if } N \leq K^2 \end{cases}
\end{aligned}
$$

Exact exercise

A first approach, that is feasible since the number of models is low, is to compute the likelihood and posterior model probabilities for each model. Using this approach yields the following posterior model probabilities:

Table: Exact Posterior Model Probabilities

Model	Probability
Only First Three	79.74%
All Others	20.26%

Model Composition

A possible second approach is to follow the algorithm proposed before and explore the possible model space. We can then compute posterior inclusion probabilities (PIP) as the amount of times a variable appears as part of a model. After 10,000 iterations and discarding 2,000, the PIP's are as follows:

Table: Posterior Inclusion Probabilities: First Method

Variable	Probability
X_1	100.00%
X_2	100.00%
X_3	100.00%
X_4	8.09%
X_5	6.20%

Model Composition

Finally, yet another approach similar to the previous one is to take M models such that $M << 2^K$ and focus on those. That is, we take an initial M models and then iterate by pairing the worse of those with a candidate model. With the same iteration scheme, the PIP's for the best $M = 5$ models are:

Table: Posterior Inclusion Probabilities: Second Method

	Probability
X_1	100.00%
X_2	100.00%
X_3	100.00%
X_4	0.00%
X_5	0.00%

References

Koop, G. (2003). *Bayesian Econometrics*. Wiley.

Madigan, D., York, J., and Allard, D. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review*, 63(2):215-232