PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1162 — 2007
Applications of Mathematics and Statistics in Economics

Dariusz Biskup

Wrocław University of Economics, Poland

BAYESIAN APPROACH TO VARIABLE SELECTION IN LINEAR REGRESSION MODEL AND ITS APPLICATION

1. Introduction

Selection of a proper set of variables in linear regression model is advisable at least for two reasons. First, the estimates of model parameters for the model with reduced number of variables tend to have smaller variance. The second reason is interpretability of the model parameters, which is possible only in the case when the model contains the variables which really influence the dependent variable. Selection of the model variables can be also seen as a method of identification of factors influencing the dependent variable.

The paper presents Bayesian approach to the variable selection (or more generally of model choice) in the linear regression model. There are several practical solutions in this approach (see i.e. [3; 4; 6]). The one used here will be the Reversible Jump Markov Chain Monte Carlo algorithm which has been proposed in [7]. The paper shows an application of this method with the usage of the so-called data prior which utilizes part of the data to produce an informative prior.

2. Regression model

Let us assume that dependent variable Y may be determined by some subset of variables belonging to the set $X_1, X_2, ..., X_k$.

The general regression equation for one of the models will have the following form:

$$Y(\mathbf{x}) = \sum_{j=1}^{r} \lambda_j \left(\left(\mathbf{x}^{(j)} \right)^T \theta^{(j)} + \varepsilon_j \left(\mathbf{x}^{(j)} \right) \right), \tag{1}$$

where:

$\boldsymbol{\theta}^{(j)} = \left[\theta_0^{(j)} \quad \theta_1^{(j)} \quad \cdots \quad \theta_{k_j}^{(j)}\right]^T$ – parameter vector for the j-th model,

k_j – number of variables selected for the j-th model,

r – number of models (maximum of which is 2^k),

$\mathbf{x} = [1 \quad x_1 \quad x_2 \quad \cdots \quad x_k]^T$ – vector of dependent variable values,

$\mathbf{x}^{(j)}$ – subset of dependent variables used in model j,

$\varepsilon_j\left(\mathbf{x}^{(j)}\right)$ – error term for model j with precision $\tau^{(j)}$ (usually assumed to be normally distributed),

λ_j – a binary parameter which is equal to one if the correct model is indexed by j.

Let M denote a discrete random variable, indicating the correct model number. This variable may take on values between 1 and r. If $M = j$, then model j $(j = 1, ..., r)$ is correct. It means that $p(M = j) = p\left(\lambda_j = 1, \lambda_{\neq j} = 0\right)$ for $j = 1, ..., r$.

The parameters λ_j are treated as random variables just like the regression parameters $\boldsymbol{\theta}^{(j)} = \left[\theta_0^{(j)} \quad \theta_1^{(j)} \quad \cdots \quad \theta_{k_j}^{(j)}\right]^T$. We will seek now the posterior distribution of the parameters λ_j which will be interpreted as model probabilities. Model with the highest probability is then considered to be correct.

The probability of model j may be computed according to the Bayes' formula in the following way:

$$p(M = j|\mathbf{X}, \mathbf{y}) = \frac{p(M = j) \cdot p(\mathbf{y}|\mathbf{X}, M = j)}{p(\mathbf{y}|\mathbf{X})}, \tag{2}$$

where:

$\mathbf{y} = [y_1 \quad \cdots \quad y_n]^T$ – vector of n values of Y,

$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$ – the matrix of n values of k dependent variables $X_1, X_2, ..., X_k$,

$p(M = j|\mathbf{X}, \mathbf{y})$ – probability of model j for data $\mathbf{X}$, $\mathbf{y}$,

$p(M = j)$ – prior probability of model j,

$p(\mathbf{y}|\mathbf{X}, M = j)$ – density of $\mathbf{y}$ if model j is correct.

Let $\vartheta^{(j)} = \left(\theta^{(j)}, \tau^{(j)}\right)$. Formula (2) can be written then in the following way:

$$p\left(M = j|\mathbf{X},\mathbf{y}\right) = \frac{p\left(M = j\right)p\left(\mathbf{y}|\mathbf{X},M = j\right)}{p\left(\mathbf{y}|\mathbf{X}\right)} = \frac{p\left(M = j\right)\int p\left(\mathbf{y},\vartheta^{(j)}|\mathbf{X},M = j\right)d\vartheta^{(j)}}{p\left(\mathbf{y}|\mathbf{X}\right)}$$

$$= \frac{p\left(M = j\right)}{p\left(\mathbf{y}|\mathbf{X}\right)}\int p\left(\mathbf{y}|\mathbf{X},M = j,\vartheta^{(j)}\right)p\left(\vartheta^{(j)}|M = j\right)d\vartheta^{(j)}.$$

If the prior probabilities $p\left(M = j\right)$ are the same for each model, then it is enough to compute the term:

$$p\left(\mathbf{y}|\mathbf{X},M = j\right) = \int p\left(\mathbf{y}|\mathbf{X},M = j,\vartheta^{(j)}\right)p\left(\vartheta^{(j)}|M = j\right)d\vartheta^{(j)}. \qquad (3)$$

The computation of (3) except for some special cases requires numerical algorithms.

3. Reversible Jump algorithm

Reversible Jump Markov Chain Monte Carlo Algorithm (RJMCMC) is a generalization of the Metropolis-Hastings algorithm for the cases when models have various numbers of parameters (i.e. dimensionality of models may differ). In cases when there is a change of dimensionality the RJMCMC iteration consists of the following steps (see [8]):

1. Let $\left(j,\vartheta^{(j)}\right)$ denote the current value of parameter vector $\vartheta^{(j)}$ of the j-th model, whose dimensionality is k_j.

2. Jump to the state $\left(j',\vartheta^{(j')}\right)$ from state $\left(j,\vartheta^{(j)}\right)$ is performed with probability $h\left(j,j'\right)$.

3. Simulate $\mathbf{u}$ from a specified distribution $q\left(\mathbf{u}|\vartheta^{(j)},j,j'\right)$.

4. Set $\left(\vartheta^{(j')},\mathbf{u}'\right) = g_{j,j'}\left(\vartheta^{(j)},\mathbf{u}\right)$, where $g_{j,j'}$ is a deterministic invertible function. This is a dimension-matching function $\left(k_j + \dim\left(\mathbf{u}\right) = k_{j'} + \dim\left(\mathbf{u}'\right)\right)$. Besides $g_{j',j} = g_{j,j'}^{-1}$.

5. Accept model change with probability $\min\left(1,A\right)$, where:

$$A = \frac{p\left(\mathbf{y}|\vartheta^{(j')},j'\right)p\left(\vartheta^{(j')}|j'\right)p\left(j'\right)h\left(j',j\right)q\left(\mathbf{u}'|\vartheta^{(j')},j',j\right)}{p\left(\mathbf{y}|\vartheta^{(j)},j\right)p\left(\vartheta^{(j)}|j\right)p\left(j\right)h\left(j,j'\right)q\left(\mathbf{u}|\vartheta^{(j)},j,j'\right)}\left|\frac{\partial g_{j,j'}\left(\vartheta^{(j)},\mathbf{u}\right)}{\partial\left(\vartheta^{(j)},\mathbf{u}\right)}\right|. \qquad (4)$$

If the dimension of model j' is higher than that of model j, then $\dim(\mathbf{u}')=0$ and function $g_{j,j'}$ performs transformation $\boldsymbol{\vartheta}^{(j')}=g_{j,j'}\left(\boldsymbol{\vartheta}^{(j)},\mathbf{u}\right)$. If dimension of model j' is lower than that of model j, then $\dim(\mathbf{u})=0$ and the transformation has the form $\left(\boldsymbol{\vartheta}^{(j')},\mathbf{u}'\right)=g_{j,j'}\left(\boldsymbol{\vartheta}^{(j)}\right)$, and value of $\mathbf{u}'$ is ignored.

If the model dimensionality is not changed, then the new values of the parameter vector may be generated for example using Gibbs algorithm.

In practice each RJMCMC iteration is usually followed by Gibbs iteration.

4. Application of RJMCMC to variable selection

4.1. Prior distributions

The following prior distributions will be assumed for model (1):

$$\boldsymbol{\theta}^{(j)} \sim N\left(\mathbf{m}^{(j)},\mathbf{V}^{(j)}\right), \tag{5}$$

$$\tau^{(j)} \sim Gamma\left(\alpha^{(j)},\beta^{(j)}\right). \tag{6}$$

$\boldsymbol{\theta}^{(j)}$ and $\tau^{(j)}$ are assumed to be independent *a priori*. It is not possible to use noninformative priors in model selection problems (see i.e. [1]). One of the possible solutions is to use a part of the data to compute informative prior and the rest for the model comparison. Other possibilities have been described in for example [1; 2; 9; 10; 12]. In the examples in paragraphs 5 and 6 randomly selected subset of the data of length equal to $\max\left\{k_j+1\right\}$ will be used.

4.2. Conditional distributions

Conditional posterior distributions of model parameters (useful when using Gibbs sampling) have the following form:

$$\boldsymbol{\theta}^{(j)}\left|\tau^{(j)}\right. \sim$$

$$N\left(\left(\tau^{(j)}\left(\mathbf{X}^{(j)}\right)^T\mathbf{X}+\left(\mathbf{V}^{(j)}\right)^{-1}\right)^{-1}\left(\tau^{(j)}\left(\mathbf{X}^{(j)}\right)^T\mathbf{y}+\left(\mathbf{V}^{(j)}\right)^{-1}\mathbf{m}^{(j)}\right),\left(\tau^{(j)}\left(\mathbf{X}^{(j)}\right)^T\mathbf{X}+\left(\mathbf{V}^{(j)}\right)^{-1}\right)^{-1}\right), \tag{7}$$

$$\tau^{(j)}\left|\boldsymbol{\theta}^{(j)}\right. \sim \Gamma\left(\alpha^{(j)}+\frac{n}{2},\beta^{(j)}+\frac{\left(\mathbf{y}-\mathbf{X}^{(j)}\boldsymbol{\theta}^{(j)}\right)^T\left(\mathbf{y}-\mathbf{X}^{(j)}\boldsymbol{\theta}^{(j)}\right)}{2}\right), \tag{8}$$

where $\mathbf{X}^{(j)}$ denotes a subset of matrix $\mathbf{X}$ containing columns appropriate for model j.

4.3. Model changing procedure

We start with a random selection of variables. After that in every iteration one of k variables is selected at random. If the selected variable is already in the current set of variables, then it is removed. Otherwise it is added to the current set. It means that $h(j',j)=h(j,j')=1/k$. We assume that the prior probabilities of each model are equal, which means that $p(j)=1/2^k$. If model j contains k variables and model j' has $k+1$ variables, then $\dim(\mathbf{u})=1$ and we will assume that $\mathbf{u}$ will be generated from the prior distribution of the added variable. The transformation function $g_{j,j'}:\left(\boldsymbol{\vartheta}^{(j)},\mathbf{u}\right)\rightarrow\boldsymbol{\vartheta}^{(j')}$ will be given by the following transformations:

$$\begin{aligned}\theta_A^{(j')}&=\theta_A^{(j)},\\ \theta_l^{(j')}&=u,\\ \tau^{(j')}&=\tau^{(j)},\end{aligned}\tag{9}$$

where A is a set of indexes for variables existing in model j and l is a number of the added variable. It means that basically the values of the parameters from the "old" model are preserved while the new parameter is generated from its prior distribution. If model j contains k variables and model j' has $k-1$ variables then $\dim(\mathbf{u}')=1$ and $\mathbf{u}'$ will be generated from the prior distribution of the deleted variable. The transformation function $g_{j,j'}:\boldsymbol{\vartheta}^{(j)}\rightarrow\left(\boldsymbol{\vartheta}^{(j')},\mathbf{u}'\right)$ will be given by the following transformations:

$$\begin{aligned}\theta_A^{(j')}&=\theta_A^{(j)},\\ \tau^{(j')}&=\tau^{(j)},\end{aligned}\tag{10}$$

where A is a set of indexes for variables existing in model j'.

The Jacobian term $\left|\dfrac{\partial g_{j,j'}\left(\boldsymbol{\vartheta}^{(j)},\mathbf{u}\right)}{\partial\left(\boldsymbol{\vartheta}^{(j)},\mathbf{u}\right)}\right|$ in formula (4) of transformation from $\left(\theta_A^{(j)},u,\tau^{(j)}\right)$ to $\left(\theta_A^{(j')},\theta_l^{(j')},\tau^{(j')}\right)$ found for functions (9) is equal to one. In the same way we find it equal to one for transformations (10).

The prior distributions $p\left(\vartheta^{(j)}|j\right)$ for formula (4) are defined as described in paragraph 4.1. The likelihood functions $p\left(\mathbf{y}\left|\vartheta^{(j)},j\right.\right)$ are found in a standard way with assumption that error term in model (1) is normally distributed.

5. Simulation example

A dataset containing 4 explanatory variables $X_1, X_2, ..., X_4$ and one dependent variable Y was simulated (the number of observations $n = 100$). Its correlation matrix is given in Table 1.

Table 1. Correlation matrix for the simulation example

	X_1	X_2	X_3	X_4	Y
X_1	1				
X_2	0.028919	1			
X_3	0.029635	0.045635	1		
X_4	0.004797	0.975174	0.032976	1	
Y	0.598903	0.817561	0.053358	0.784662	1

As we see, there are three explanatory variables highly correlated to Y. These are: X_1, X_2, X_4. However, there is also strong correlation between X_2 and X_4. After performing 100 000 RJMCMC simulations, the probability of the correct model which contains only variables X_1 and X_2 amounted to 0.99842. The model with variables X_1, X_2, X_4 has probability 0.00158.

6. GDP Prediction Example

The dataset used in the example has been taken from [5]. The predicted variable is GDP in 1992. In the original study there were 41 regressors. Here only 8 of them have been chosen. The following variables have been used:

X_1 – country area,

X_2 – primary school enrollment,

X_3 – life expectancy,

X_4 – GDP level in 1960,

X_5 – no. of years of open economy,

X_6 – % of English speakers,

X_7 – years of capitalism,

X_8 – equipment investment.

The correlation matrix is given in Table 2.

Table 2. Correlation matrix for the GDP example

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	1								
X_1	–0.148662	1							
X_2	0.448613	–0.147381	1						
X_3	0.491026	–0.287173	0.829518	1					
X_4	0.211511	–0.326806	0.737224	0.869197	1				
X_5	0.370339	–0.191349	0.488538	0.477146	0.448789	1			
X_6	0.602564	–0.357154	0.605759	0.706425	0.63405	0.521083	1		
X_7	0.013973	0.023219	–0.265304	–0.30613	–0.230955	–0.03293	–0.19523	1	
X_8	–0.362621	0.03982	–0.480395	–0.450342	–0.379677	–0.25405	–0.38809	0.324746	1

Model probabilities have been computed using 100 000 iterations of the RJMCMC algorithm. The largest have been shown in Table 3.

Table 3. Probability distribution across selected models

Variable set	Probability
X_3, X_4, X_6	0.01798
X_2, X_3, X_4, X_6	0.01086
X_3, X_4, X_6, X_7	0.40811
X_1, X_3, X_4, X_6, X_7	0.06414
X_2, X_3, X_4, X_6, X_7	0.21786
$X_1, X_2, X_3, X_4, X_6, X_7$	0.04029
$X_2, X_3, X_4, X_5, X_6, X_7$	0.03226
X_3, X_4, X_5, X_7, X_8	0.01872
X_3, X_4, X_6, X_7, X_8	0.07918
$X_2, X_3, X_4, X_6, X_7, X_8$	0.02259
X_3, X_4, X_5, X_6	0.02179

We can see that the most likely set of variables is X_3, X_4, X_6, X_7. The other one with significant probability is X_2, X_3, X_4, X_6, X_7 which differs from the first one through addition of the X_2 variable.

In cases where there is no model with significantly highest probability, it may be advisable to use the so called model-averaging (see [11]). We do not choose

then the one best model, but make predictions for the dependent variable from all the models and weigh them using the corresponding model probability.

References

[1] Aitkin M., "Posterior Bayes Factors", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1991, 53, 111-142.

[2] Berger J., Pericchi L., "The Intrinsic Bayes Factor for Model Selection and Prediction", *Journal of the American Statistical Association* 1996, 91, 109-122.

[3] Carlin B., Chib S., "Bayesian Model Choice via Markov Chain Monte Carlo Methods", *Journal of the Royal Statistical Society, Series B* 1995, 75 (3), 473-484.

[4] Dellaportas P., Forster J.J., Ntzoufras I., "On Bayesian Model and Variable Selection Using MCMC", *Statistics and Computing* 2002, 12, 27-36.

[5] Fernandez C., Ley E., Steel M., "Model Uncertainty in Cross-country Growth Regressions", *Journal of Applied Econometrics* 2001, 16, 563–576.

[6] George E., McCulloch R., "Variable Selection via Gibbs Sampling", *Journal of the American Statistical Association* 1993, 88 (423), 881-889.

[7] Green P., "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination", *Biometrika* 1995, 82, 711–732.

[8] Han C., Carlin B.P., "MCMC Methods for Computing Bayes Factors: A Comparative Review", *Journal of the American Statistical Association* 2001, 96, 1122-1132.

[9] O'Hagan A., "Fractional Bayes Factors for Model Comparison", *Journal of the Royal Statistical Society, Series B* 1995, 57, 99-138.

[10] Perez J.M., Berger J., "Expected Posterior Prior Distributions for Model Selection", *Biometrika* 2002, 89, 491-512.

[11] Raftery E.A., Madigan D., Hoeting A., "Bayesian Model Averaging for Linear Models", *Journal of the American Statistical Association* 1997, 92, 179-191.

[12] Spiegelhalter D.J., Smith A.F.M., "Bayes Factors for Linear and Log-Linear Models with Vague Prior Information", *Journal of the Royal Statistical Society, Series B* 1982, 44, 377-387.

PODEJŚCIE BAYESOWSKIE DO WYBORU ZMIENNYCH W MODELU REGRESJI ORAZ JEGO ZASTOSOWANIE

Streszczenie

W artykule ukazano bayesowskie podejście do problemu doboru zmiennych w modelu regresji liniowej. W tym podejściu dobór zbioru zmiennych dokonuje się przez poszukiwanie modelu o największym prawdopodobieństwie zaistnienia. Ponieważ analityczne obliczenie tego prawdopodobieństwa jest w większości przypadków niemożliwe, została wykorzystana metoda *reversible jump*. Metoda ta należy do klasy algorytmów typu MCMC (*Markov Chain Monte Carlo*) przystosowanych do przestrzeni o zmiennej liczbie wymiarów. W artykule przedstawiony jest przykład symulacyjny ze współliniowymi zmiennymi, a także przykład z rzeczywistymi danymi dotyczący predykcji PKB.

Słowa kluczowe: *Markov Chain Monte Carlo*, regresja liniowa, wybór modelu, algorytm *reversible jump*.