基于 Gibbs Sampler 的线性回归模型选择

赵昕东，耿　鹏

（华侨大学 数量经济研究院，福建 泉州 362021）

摘要：线性回归模型是计量经济学的基本模型，在建立线性回归模型的过程中，模型选择是非常重要的一个环节，如果可能的解释变量不是很多时，可以通过逐步回归的方法比较每个候选模型的准则值，如 AIC、SIC 等进行模型选择。可是，当存在大量可能的解释变量时，我们无法一一比较每个候选模型的准则值。为了解决这个问题，文章提出一个基于 Gibbs Sampler 的线性回归模型选择方法，结果表明应用该方法能够从大量候选模型中准确、高效地确认准则值最小的模型。

关键词：线性回归模型；模型选择；Gibbs Sampler；准则值

中图分类号：F224.0　　**文献标识码**：A　　**文章编号**：1001－5124（2009）04－0089－05

一、前言

线性回归模型是计量经济学中最基本的模型，应用范围非常广泛。在实际应用中，经常会遇到在建立模型时大量可能的解释变量的取舍问题。常用的模型检验方法是逐步回归方法，即通过 t 统计量检验单个变量的显著性，通过 F 统计量检验整个方程的显著性，并逐步剔除不显著的解释变量。此外还有准则值方法，即通过比较 AIC 和 SIC 等准则函数值确定准则值最小的模型为最优模型。Clayton、Geisser 和 Jennings 指出准则函数方法比其他方法更有效，[1] Granger、King 和 White 指出该方法受到的限制比其他方法少。[2]但是在应用逐步回归方法寻找准则值最小的模型时，当解释变量数量较大时，逐步回归方法会浪费大量时间，缺乏效率并且不能保证找到最优模型。例如，存在 15 个可供选择的解释变量时，不考虑常数项仍将存在 $2^{15}=32768$ 个候选模型，一一计算这些模型的准则值，然后进行比较是不现实的。本文中我们建立了一种利用 Gibbs Sampler 的模型选择方法，其思想是：将候选模型的准则值与生成该模型的概率相联系，使准则值最小的模型拥有最大的概率，由全部解释变量的联合分布通过 Gibbs Sampler 生成随机模型，那么随机模型的样本中最优模型即最小准则值模型出现的次数最多。

基于 Gibbs Sampler 的模型选择方法是一种基于 AIC、SIC 等准则函数方法进行模型选择时遇到大量候选模型情况下的一种解决方法，在进行模型选择时仍然要通过比较不同候选模型的准则函数值确定最优模型，应用我们的方法的目的是在应用准则函数方法时可以快速高效地确定最优模型。

本文首先介绍 Gibbs Sampler，然后介绍线性回归模型及模型选择的一些概念，最后介绍如何应用 Gibbs Sampler 进行线性回归模型选择，最后我们将这种方法应用到模拟数据上用来验证该方法的有效性。

二、Gibbs Sampler

吉伯斯样本生成器最早由 Geman 和 Geman 年提出，[3]具体算法如下：假定 $X=(X_1,X_2,\cdots X_k)$ 是 K 维随机变量，其联合分布是 f，而且 $X_1,X_2,\cdots X_k$ 的条件分布分别是 $f_1,f_2,\cdots f_k$。吉伯斯样本生成

收稿日期：2008－12－05

基金项目：教育部人文社会科学一般项目（07JA790004）；福建省自然科学基金（2009J01312）；华侨大学科研基金（07BS501）。

第一作者简介：赵昕东（1968－），男，吉林长春人，华侨大学数量经济研究院研究员，博士生导师。

器就是在给定 $x^{(t-1)}=(x_1^{(t-1)},x_2^{(t-1)},\cdots x_k^{(t-1)})$ 条件下，根据下面的条件分布生成 $X_1,X_2,\cdots X_k$ 的随机样本 $x^{(t)}=(x_1^{(t)},x_2^{(t)},\cdots x_k^{(t)})$ 的方法。

$$X_1^{(t)} \sim f_1(x_1 \mid x_2^{(t-1)},\cdots,x_k^{(t-1)})$$
$$X_2^{(t)} \sim f_2(x_2 \mid x_1^{(t)},x_3^{(t-1)},\cdots,x_k^{(t-1)})$$
$$\vdots$$
$$X_i^{(t)} \sim f_i(x_i \mid x_1^{(t)},\cdots,x_{i-1}^{(t)},x_{i+1}^{(t-1)}\cdots,x_k^{(t-1)})$$
$$\vdots$$
$$X_k^{(t)} \sim f_k(x_k \mid x_1^{(t)},\cdots,x_{k-1}^{(t)})$$

吉伯斯样本生成器的优点是通过多元分布的各个边际分布生成随机样本，这些样本可以看成是多元分布的随机样本，从而将多元分布的蒙特卡洛计算简化成一元的蒙特卡洛计算，避免了直接通过多元分布生成样本可能带来的困难。

三、线性回归模型选择

常用的线性回归模型选择方法是逐步回归，即通过变量的显著性检验（t 检验），拟合优度检验（F 检验），逐步剔除不显著的解释变量，但是逐步回归不能保证最终发现最优模型。通过比较所有候选模型的准则函数值可以确保找到该准则函数意义上的最优模型，常用的准则函数有 AIC 准则和 SIC 准则：

$$AIC=-2l/T+2(k+1)/T$$
$$SIC=-2l/T+(k+1)\ln T/T$$

其中 l 为对数似然值，T 为样本长度，k 为解释变量个数。

但是应用准则函数方法可能遇到一个难题，那就是，当解释变量数量较多时（如 15 个），则候选模型数量为 2^{15}，这种情况下，我们无法一一比较每个候选模型的准则值。在实际问题中这种情况经常会发生，如果某个被解释变量有 5 个影响因素，如果考虑这 5 个影响因素 1 阶滞后与 2 阶滞后，则存在 15 个可能的解释变量。为解决这一问题，本文提出了一个应用 Gibbs Sampler 的模型选择方法，应用这个方法，可以快速准确的找到准则值最小的模型。

四、Gibbs Sampler 与线性回归模型选择

假定我们已经确定了线性模型的所有可能的 n 个解释变量 $x_1,x_2,\cdots x_n$，则共有 2^n 个候选模型，模型选择就是确定哪些解释变量的系数为 0。我们定义一个向量 $V=(v_1,v_2,\cdots v_n)$，其中 v_i 为 0 或 1，v_i 为 0 表明 x_i 的系数为 0，v_i 为 1 表明 x_i 的系数不为 0。则 V 可以用来代表某个候选模型的结构。

用 MC 代表任意一种准则函数，则 AIC 与 BIC 可以统一表示成：

$$MC(V)=\log\hat{\Sigma}+C(N)$$

V 可以看成是一个随机变量，我们定义 M 上随机变量 V 的概率分布：

$$P_\lambda(V)=\frac{\exp(-\lambda MC(V))}{\sum_{V\in M}\exp(-\lambda MC(V))}$$

这里 $0<\lambda<1$ 是起调节作用的参数，通过调节 λ 在应用 Gibbs Sampler 算法时可以增大最优模型出现的概率。这里 V 的可能取值是所有 2^n 个候选模型，每个候选模型对应一个概率值，根据定义最优模型对应最大概率值。V 的每个元素 v_i 可以取值 0 或 1，因此随机变量 v_i 服从伯奴里分布，而 V 服从 2^n 个伯奴里分布组成的联合分布，并且 V 的任意一个元素 v_i 的条件分布是：

$$\Pr\{v_i=1|v_{-i}\}=\frac{\dfrac{\exp\{-\lambda MC(V)|_{V_i=1}\}}{\sum_{V\in M}\exp\{-\lambda MC(V)\}}}{\dfrac{\exp\{-\lambda MC(V)|_{V_i=1}\}}{\sum_{V\in M}\exp\{-\lambda MC(V)\}}+\dfrac{\exp\{-\lambda MC(V)|_{V_i=0}\}}{\sum_{V\in M}\exp\{-\lambda MC(V)\}}}=\frac{1}{1+\exp\{\lambda MC(V)|_{V_i=1}-\lambda MC(V)|_{V_i=0}\}}$$

$\Pr\{v_i = 0 \mid v_{-i}\} = 1 - \Pr\{v_i = 1 \mid v_{-i}\}$

这里 v_{-i} 代表 V 中 v_i 以外的所有元素。

有了 v_i 的条件分布，我们就可以应用 Gibbs Sampler 生成服从 $P_\lambda(V)$ 的随机样本，即随机模型。具体算法如下：

步骤 1：任意选取初值 $V^{(0)}$，例如 $V^{(0)} = \{1\}_{1\times n}$。

步骤 2：在已经生成 $V^{(1)},\cdots,V^{(k-1)}$ 的条件下，按照以下步骤生成 $V^{(k)}$

对 $i = 1,2\cdots,n$ 循环

根据 $v_i^{(h)}$ 取 1 的概率 $\Pr\{v_i^{(h)} = 1 \mid v_{-i}^{(h)}\}$ 生成随机样本 1 或 0，并更新 $v_i^{(h)}$。

步骤 3：重复步骤 2，直到生成 $V^{(h+1)},\cdots,V^{(H)}$

当我们生成一定数量的随机模型后，如何确定最优模型？因为最优模型在 M 上的概率最高，因此当我们应用吉伯斯样本生成器生成随即模型时，最优模型将趋于较早出现，并且当生成的模型数量足够多时，最优模型将趋于以最大的频率出现。因此当生成足够多的模型样本时，我们可以将样本中准则值最小的模型确定为最优模型，也可以将出现频率最高的模型确定为最优模型。

五、模拟实验

为了验证该方法的有效性，我们进行一组模拟检验。本文所用的计算程序都由 S-plus 8.0 软件编写，包括数据的生成，线性模型的估计，Gibbs Sampler 算法等。

首先，我们生成样本容量为 500 的两两互不相关的平稳序列 $X_1\cdots X_{15}$ 和服从标准正态分布的随机变量 ε，令

$y = 50 + 3x_1 + 5x_2 + 4x_3 - 3x_4 + 2x_5 - x_6 + \varepsilon$

我们假设 $X_1\cdots X_{15}$ 为可能的解释变量，目标是从这 15 个可能的解释变量中确定 $X_1\cdots X_6$。加上常数项，存在 $2^{16} = 65536$ 个候选模型，显然无法一一比较每个模型的准则值。以下我们应用本文提出的 Gibbs Sampler 模型选择方法进行模型选择。

λ 分别取 0.4、0.6 和 0,8，分别生成 100 和 200 个随机模型，结果见表 1。

表 1　模型选择结果

	生成随机模型个数	出现最多模型的次数	出现最多模型的频率	出现最多模型的 SIC 值	出现最多模型的结构
λ=0.4	100	54	54%	3166.228	1111111000000000
	200	90	45%	3166.228	1111111000000000
λ=0.6	100	80	80%	3166.228	1111111000000000
	200	154	77%	3166.228	1111111000000000
λ=0.8	100	80	80%	3166.228	1111111000000000
	200	170	85%	3166.228	1111111000000000

从表 1 可以看到在所有各种情况下，真实模型（1111111000000000）出现的次数都是最多的，出现的频率也是最高的。如果生成 100 个随机模型，则只需进行 100*14*2=2800 次模型估计，与 65536 相比，大大提高了效率。

最后，我们利用选择的模型结构进行估计，结果如表 2：

表 2　模拟模型估计结果

截距项	x_1	x_2	x_3	x_4	x_5	x_6
50.026	2.999	5.008	3.999	−2.998	1.999	−1.005

六、实际应用

我们将 Gibbs Sampler 运用到实际数据中，利用 Goldfeld[4]建立的美国货币需求估计的局部校正模型 $\ln m_t = b_0 + b_1 \ln y_t + b_2 r_t + b_3 \ln m_{t-1} + b_4 \pi_t + \varepsilon_t$ 来验证中国的货币现象。在 Goldfeld 和 Sichel 的货币需求模型中，m_t 表示货币供给，解释变量中 y_t 表示产出，r_t 表示短期利率，π_t 表示通货膨胀率。我们

在 Goldfeld 和 Sichel 的模型基础上加入可能影响货币供给的法定准备金率 e_t，将解释变量的 1 至 3 阶滞后值也作为可能的解释变量，这样包括常数项共有 20 个可能的解释变量，分别是截矩项 b_0、y_t、y_{t-1}、y_{t-2}、y_{t-3}、r_t、r_{t-1}、r_{t-2}、r_{t-3}、m_{t-1}、m_{t-2}、m_{t-3}、π_t、π_{t-1}、π_{t-2}、π_{t-3}、e_t、e_{t-1}、e_{t-2}和 e_{t-3}。

本文中用名义货币 M2 反映货币需求 m_t，用实际季度 GDP 反映产出 y_t，用 CPI 同比增长率反映通货膨胀率 π_t，用银行同业间拆借利率（7 天期）反映短期利率 r_t。数据来自于中国国家统计局和中国人民银行（月度数据均取第 3，6，9，12 月转换为季度数据），时间跨度为 2001 年 3 月到 2007 年 12 月，其中 GDP 和 M2 数据均利用 X11 季节调整方法去掉了季节因素。

以 SIC 为准则函数，$\lambda = 0.8$，生成 300 个随机模型，计算结果显示模型（00000100010010000000）出现次数最多，共出现 59 次，出现频率为 19.7%，其余候选模型出现频率均未超过 10%，可以认为此模型为最优模型。最优模型估计结果见表 3。

表 3　货币需求函数初步估计结果

	参数估计值值	t 值
b_0	1.5191	4.1685
$\ln \pi_t$	−0.3362	−3.8028
$\ln m_{t-1}$	1.0055	186.2781
r_t	0.0110	1.9632

由于滞后的因变量 X8 的系数接近 1，说明出现了错误设定的局部校正，而且其他变量的符号也与理论不符，出现了同戈德费尔德在估计 1974 年以后美国经济数据时相同的问题，因此我们对该模型进行修改，去除其局部校正机制，即从解释变量中去掉被解释变量的滞后值，这样包括常数项共有 17 个可能的解释变量，分别是截矩项 b_0、y_t、y_{t-1}、y_{t-2}、y_{t-3}、r_t、r_{t-1}、r_{t-2}、r_{t-3}、π_t、π_{t-1}、π_{t-2}、π_{t-3}、e_t、e_{t-1}、e_{t-2}和 e_{t-3}

以 SIC 为准则函数，$\lambda = 0.8$，生成 300 个随机模型，计算结果显示模型（01100100011011000）出现次数最多，共出现 81 次，出现概率为 27%，其余备选模型出现概率均未超过 10%，因此认为该模型为最优模型。最优模型估计结果见表 4：

表 4　货币需求函数最终估计结果

	参数估计值值	t 值
b_0	−5.7309	−3.9708
$\ln y_t$	0.7121	8.9982
$\ln y_{t-1}$	0.3148	4.2067
r_t	0.0621	2.8356
π_t	1.0699	3.0334
π_{t-1}	2.9025	7.0399
π_{t-3}	−2.2191	−8.0696
e_{t-1}	−0.4143	−6.8717

从这个模型我们可以看出来，除名义利率外，其他变量的符号都符合经济理论，这可能是由于中国货币市场不完善，货币需求对利率不够敏感造成的。

七、结论

本文中我们提出了利用 Gibbs Sampler 的线性模型选择方法，从实验结果表明该方法具有较好的使用效果，能够节省大量的运算时间，并且通过在实际问题上的应用，说明该方法也具有可行性。

参考文献

[1] CLAYTON M K, GEISSER S, JENNINGS D E. A comparison of several model selection procedures [M]//GOEL P, ZELLNER A. (eds.), Bayesian Inference and Decision Techniques. New York: Elsevier Science Publisher, 1986：199−212.

[2] GRANGER C J W, KING M L, WHITE H. Comments on the testing economic theries and the use of model selection criteria [J]. Journal of

Econometrics, 1995, 67：173–187.

[3] GEMAN S, GEMAN D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 6：721–741.

[4] GOLDFELD S M. "The Demand for Money Revisited"[J]. Brookings Papers on Economic Activity, 1973, 3：577–638.

Linear Regression Model Selection Based on Gibbs Sampler

ZHAO Xin–dong, GENG Peng

(*Institute for Quantitative Economics, Huaqiao University, Quanzhou* 362021, *China*)

Abstract: The Linear Regression Model is the basic model in Econometrics. When establishing the model, model selection is a very important task. When the number of potential explanatory variables is not too large, one can perform model selection by comparing the criterion values, such as AIC, SIC of each model. However, when there are a large number of potential explanatory variables, it is impossible to compare the criterion values of the candidate models one by one. To solve this problem, a model selection procedure based on Gibbs Sampler will be proposed. The results show that such a method enables us to identify the model with the smallest criterion value accurately and efficiently among a a large number of candidate models.

Key Words: Linear Regression Model; model selection; Gibbs Sampler; criterion value

（责任编辑　王　抒）

＊＊＊

（上接第 83 页）

[7] 马克思，恩格斯. 马克思恩格斯选集：第 1 卷[M]. 北京：人民出版社，1995.

[8] 梁启超. 梁启超选集[M]. 上海：上海人民出版社，1984.

[9] 马克思，恩格斯. 马克思恩格斯选集：第 3 卷[M]. 北京：人民出版社，1995.

[10] 毛泽东. 毛泽东选集：第 2 卷[M]. 北京：人民出版社，1991.

[11] 沈卫威. 自由守望——胡适派文人引论[M]. 上海：上海文艺出版社，1997.

[12] 许纪霖. 许纪霖自选集[M]. 桂林：广西师范大学出版社，1999.

[13] 马克思，恩格斯. 马克思恩格斯选集：第 4 卷[M]. 北京：人民出版社，1995.

The relationship between Liberalism and Marxism in Modern China and its Impact

SONG Xiao–min

(*Political and Administrative Institute, Liaoning Normal University, Dalian* 116029, *China*)

Abstract: Liberalism, as the mainstream thought of western society, has a long history. It is the important theory to which the development of capitalist civilization is pegged and one of the basic ideas of modern civilization. Western liberalism has had a critical impact on the Chinese society after it was introduced into modern China. There is an inextricably close relationship between liberalism and Marxism. Modern China's liberalism and Marxism, both originating from the West, play an equally important role in China's social development. Based on this, this paper elaborates the different fractions of Modern China's liberalism, their relationship with Marxism, as well as their different impacts on Chinese society.

Key Words: modern; China; liberalism; Marxism

（责任编辑　王　抒）