

基于贝叶斯统计思想实现多重线性回归分析

谷恒明¹, 胡良平^{1,2*}

(1. 军事医学科学院生物医学统计学咨询中心, 北京 100850;

2. 世界中医药学会联合会临床科研统计学专业委员会, 北京 100029

* 通信作者: 胡良平, E-mail: lphu812@sina.com)

【摘要】 本文目的是介绍基于贝叶斯统计思想实现多重线性回归分析的方法。多重线性回归分析时, 单纯基于贝叶斯理论导出的公式来估计回归模型中参数的做法并不常见。最常见的做法是基于马尔科夫链蒙特卡罗 (MCMC) 方法来实现多重回归分析, 即把蒙特卡罗方法、贝叶斯统计思想和马尔科夫链等内容有机结合起来, 共同完成多重回归分析。在资料基本满足经典统计思想建模的前提条件时, 基于贝叶斯统计思想构建多重线性回归模型, 其效果等价于基于经典统计思想构建的多重线性回归模型。

【关键词】 随机模拟; Metropolis 算法; Gibbs 抽样; MCMC; 先验分布; 后验分布

中图分类号: R195.1

文献标识码: A

doi:10.11886/j.issn.1007-3256.2018.01.003

Realization of a multiple linear regression analysis based on Bayesian statistical thought

Gu Hengming¹, Hu Liangping^{1,2*}

(1. Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing 100850, China;

2. Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing 100029, China

* Corresponding author: Hu Liangping, E-mail: lphu812@sina.com)

【Abstract】 The paper introduced a method based on Bayesian statistics for realising a multiple linear regression analysis. Instead of estimating the parameters in a regression model based on the formulas derived from Bayesian theory, the most common approach was based on the MCMC method to achieve the multiple regression analysis. The MCMC method was assembled Bayesian theory, Markov chain, Monte Carlo methods and other contents. When the data basically met the prerequisites of a classical statistical modeling, constructing the multiple linear regression models based on Bayesian statistical thought was equivalent to the one based on the classical statistical method.

【Keywords】 Random simulation; Metropolis algorithm; Gibbs sampling; Markov chain Monte Carlo; Prior distribution; Posterior distribution

1 MCMC 方法概述

1.1 基于贝叶斯统计思想建立回归模型的基本思路

若单纯基于贝叶斯统计思想来建立多重线性回归模型, 需要进行“共轭先验下的贝叶斯推断”和(或)“广义先验下的贝叶斯推断”求出多重线性回归模型中的“估计回归系数矩阵”和“估计误差协方差矩阵”^[1]。然而, 随着统计学的发展, 统计学家又提出了新的思路: 基于随机抽样进行统计模拟并结合贝叶斯统计思想进行计算, 能够得到基于数万次、甚至数十万次抽样样本条件下更加稳定的回归系数估计结果, 误差很小, 结果的精确度很高。这个方法叫做马尔科夫链蒙特卡罗方法, 简称为 MCMC (Markov chain Monte Carlo) 方法^[2-3]。

1.2 马尔科夫链

马尔科夫链是一种具有马尔科夫性的特别随机

过程。所谓随机过程, 用通俗的语言来表述, 就是以“时间”为自变量的一个变量(即时间的函数)。例如我国自 2000 年以来各年某种疾病的发病率资料, 每个发病率数据都与特定的年份有关。这样的一串数被称为“时间序列”, 也被称为一个随机过程。

那么, 什么样的随机过程可以被称为马尔科夫链呢? 用通俗的语言来表述, 在随机过程或时间序列中, 任何一个时间点的取值仅受其前一个时间点上取值的影响, 而不受其他时间点上的数值影响, 这一特性被称为“马尔科夫性”, 并且, 人们称具有马尔科夫性的随机过程或时间序列为“马尔科夫链”。

1.3 蒙特卡罗方法的基本思路与实质

蒙特卡罗是摩纳哥国的世界闻名赌城, 其实, 蒙特卡罗方法与赌城“蒙特卡罗”之间没有任何关联! 事实上, 美国在研究原子弹的过程中, 遇到了复杂的有关核反应的计算问题, 波兰裔美国数学家斯塔尼斯拉夫·吴拉姆 (Stanislaw Ulam) 创造性地提出了随机模拟方法, 用来计算当时遇到的复杂计算问题,

项目基金: 国家高技术研究发展计划课题资助 (2015AA020102)

并由科学家冯·纽曼(Von Neumann)在计算机上实现。为了保密,他们就将该方法称为“蒙特卡罗(Monte Carlo)”法^[3-4]。

蒙特卡罗方法的基本思想是模拟从总体中反复抽取一批又一批样本,然后利用所抽取的多批模拟样本进行估计、假设检验等统计推断。所以,蒙特卡罗方法的实质就是基于有限的样本资料,通过反复抽取随机样本,使“原先的样本”接近其所对应的“总体”情况。换一句话说,蒙特卡罗方法的实质是基于“有限样本所提供的信息”去洞察“总体的真实情况”。

1.4 马尔科夫链蒙特卡罗方法

前面介绍蒙特卡罗方法时提到了一个关键词,即“随机抽样”。随机抽样不等于“胡乱”或“随便”操作,而是依据一定的“规则”实施抽样,这里所说的“规则”在数学上就叫做“抽样算法”。

非常著名的抽样算法有:梅切波利斯(Metropolis, 1953)算法、由哈斯廷斯(Hastings, 1970)对梅切波利斯算法改进的算法(即MH算法)、吉曼(Geman, 1984)兄弟在研究数字图像恢复问题时提出的吉布斯抽样法(Gibbs sampling)以及其他一些模拟退火方法(simulated annealing)^[5],现在人们将所有这些方法和各种各样的推广统称为马尔科夫链蒙特卡罗(MCMC)方法。

2 实例及基于贝叶斯统计思想实现多重线性回归分析

2.1 问题与数据

【例1】26例糖尿病患者的血清总胆固醇(X_1)、甘油三酯(X_2)、空腹胰岛素(X_3)、糖化血红蛋白(X_4)、空腹血糖(Y)的测量值列于表1,试基于贝叶斯统计思想建立血糖与其他几项指标的多重线性回归方程。

2.2 回归分析任务

对例1表1中的数据,设因变量为 Y ,自变量为 X_1 、 X_2 、 X_3 、 X_4 ,试建立因变量依赖自变量的多重线性回归模型,并做相应的假设检验。

表 1 26 例糖尿病患者血样中有关指标的测定结果					
i	X_1	X_2	X_3	X_4	Y
1	5.68	1.9	4.53	8.2	11.2
2	3.97	1.64	7.32	6.9	8.8
...	...	...	...	...	...
25	11.54	10.89	1.2	10.5	20
26	3.84	1.2	6.54	9.6	10.4

注:详细数据见本期第一篇文章《多重线性回归分析的核心内容与关键技术概述》

2.3 采用经典统计思想筛选自变量

实际资料中往往包含了很多自变量,而其中有些自变量对因变量的影响作用无统计学意义,故需要对自变量进行筛选;在SAS中,并非所有用于回归分析的SAS过程都具有筛选自变量的功能。在处理多重线性问题时,一般都用REG过程来实现对自变量的筛选。这部分内容参见本期第二篇文章《基于经典统计思想实现多重线性回归分析》,对例1资料筛选的结果为:保留 X_2 、 X_3 、 X_4 三个自变量。

2.4 采用贝叶斯回归分析方法建模——指定先验分布为 $N(0, 10^6)$

(1)基于自变量筛选结果并借助SAS中MCMC过程创建回归模型所需要的SAS程序

```
proc mcmc data = cra1 seed = 12345 nbi = 10000
nmc = 100000 thin = 10 diag = all OUTPOST = POST
STATS = ALL MCHISTORY = DETAILED dic;
parms beta0 0 beta1 100 beta2 0 beta3 10 ;
parms sigma2 20;
prior beta0 beta1 beta2 beta3 ~ normal(mean = 0, var = 1e6);
prior sigma2 ~ igamma(shape = 0.001, scale = 0.001);
mu = beta0 + beta1 * x2 + beta2 * x3 + beta3 * x4;
model y ~ normal(mu, var = sigma2);
run;
```

【说明】MCMC过程中的语句和选项的含义比较复杂,因篇幅所限,此处从略。

(2)多重线性回归分析的输出结果

Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
beta0	10000	4.9274	2.2398	3.4456	4.9053	6.3989
beta1	10000	0.4389	0.1417	0.3454	0.4388	0.5333
beta2	10000	-0.3007	0.1009	-0.3666	-0.3000	-0.2346
beta3	10000	0.8120	0.2143	0.6744	0.8103	0.9547
sigma2	10000	3.2286	1.0766	2.4805	3.0266	3.7333

多重线性回归方程为：
 $\hat{Y}=4.9274+0.4389X_2-0.3007X_3+0.8120X_4$
相对误差绝对值的平均值为 0.1149813505，残差平方和 64.4418，决定系数 $R^2=0.70787$ 。

2.5 采用贝叶斯回归分析方法建模——无信息先验分布

(1) 基于自变量筛选结果并借助 SAS 中 MCMC 过程创建回归模型所需要的 SAS 程序
proc mcmc data =cra1 seed =12345 nbi =10000
nmc =100000 thin =10 diag = all OUTPOST = POST

STATS = ALL MCHISTORY = DETAILED dic;
parms beta0 0 beta1 100 beta2 0 beta3 10;
parms sigma2 20;
prior beta0 beta1 beta2 beta3 ~ general(0);
priorsigma2 = -log(sigma2);
prior sigma2 ~ general(priorsigma2);
mu = beta0+ beta1 * x2 + beta2 * x3 +beta3 * x4;
model y ~ normal(mu,var = sigma2);
run;

(2) 多重线性回归分析的输出结果

Posterior Summaries						
Parameter	N	Mean	StandardDeviation	Percentiles		
				25%	50%	75%
beta0	10000	4.9222	2.2498	3.4316	4.9064	6.3549
beta1	10000	0.4361	0.1406	0.3443	0.4359	0.5278
beta2	10000	-0.3015	0.1012	-0.3688	-0.3009	-0.2349
beta3	10000	0.8135	0.2150	0.6749	0.8165	0.9525
sigma2	10000	3.1965	1.0689	2.4537	2.9872	3.7111

多重线性回归方程为：
 $\hat{Y}=4.9222+0.4361X_2-0.3015X_3+0.8135X_4$
相对误差绝对值的平均值为 0.11495609，残差平方和 64.4434，决定系数 $R^2=0.70786$ ，标准化均方误差 NMSE = $1-R^2=0.292140$ 。

基于经典统计思想和贝叶斯统计思想(给定两种先验分布)建立多重线性回归分析模型,对模型拟合效果进行评价的结果见表 2。

表 2 基于经典统计与贝叶斯统计建模模型评价结果的比较

模型评价指标	模型评价指标具体值		
	传统统计	指定先验分布	无信息先验
相对误差绝对值均值	0.1149238927	0.1149813505	0.11495609
残差平方和	64.4411	64.4418	64.4434
决定系数	0.70787	0.70787	0.70786
NMSE	0.29213	0.29213	0.29214

注:先验分布指正态分布、无信息先验实际上是取均匀分布作为先验分布;NMSE 为标准化均方误差

3 小 结

数据本身质量(即自变量间不存在多重共线性关系、无严重的异常点)较好时,即经典统计思想建模方法所要求的假定条件基本上能够得到满足,此时,经典统计建模的效果就比较好;如果有足够准确的先验信息,采用经典统计思想进行多重线性回归分析和采用贝叶斯统计思想回归分析的结果相差不大。

【说明】感兴趣的读者可以参照本期第二篇文章《基于经典统计思想实现多重线性回归分析》中的做法,产生派生变量,并基于自变量筛选得到 Z_4 、 Z_5 、 Z_6 三项,再将这三项分别代入本文中两段调用 MCMC 过程的 SAS 程序中去计算,会得到相应的多重线性回归分析结果。因篇幅所限,此处从略。

值得一提的是:文献[6-9]中的资料都可以采用本文介绍的方法进行多重线性回归分析,其结果和结论更具有可信性。

参考文献

[1] 刘金山,夏强. 基于 MCMC 算法的贝叶斯统计方法[M]. 北京: 科学出版社, 2017: 118-174.
[2] 茆诗松,王静龙,濮晓龙. 高等数理统计[M]. 北京: 高等教育出版社, 2010: 440-459.
[3] 黄长全. 贝叶斯统计及其 R 实现[M]. 北京: 清华大学出版社, 2017: 114-138.
[4] 高惠璇. 统计计算[M]. 北京: 北京大学出版社, 2017: 173-223.
[5] 康崇禄. 蒙特卡罗方法理论和应用[M]. 北京: 科学出版社, 2017: 86-440.
[6] 任传波,姜季妍,董黎明. 青少年抑郁障碍患者心理社会特征[J]. 四川精神卫生, 2017, 30(5): 455-457.
[7] 徐华丽,孙崇勇,高悦. 大学生人格特质对手机成瘾倾向的影响[J]. 四川精神卫生, 2017, 30(5): 458-462.
[8] 李青青,张倩. 医学院校大学生情绪状态与学业拖延的关系[J]. 四川精神卫生, 2017, 30(5): 463-465.
[9] 魏国英,曾丽娟,周桂成,等. 精神科护士职业倦怠与工作压力的相关性[J]. 四川精神卫生, 2017, 30(5): 466-469.

(收稿日期:2018-01-29)
(本文编辑:陈霞)