Rstan在Bayesian分析中的运用

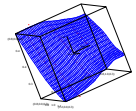
李政宵

统计学院, 风险管理与精算学
中国人民大学, 北京
li_zhengxiao@126.com

主要模型

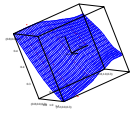
数据和代码 (https://github.com/lzx89757/Introduction-to-Rstan)

1. Rstan 基本介绍 ✓
2. 基本分布拟合
3. 线性回归模型
4. 广义线性模型
5. 案例分析-准备金数据分析
6. 总结



Stan: Help

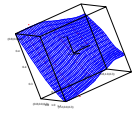
- Installing: https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Windows
- Homepage: http://mc-stan.org/interfaces/rstan
- Example Models
  https://github.com/stan-dev/example-models/wiki
- Stan Users group:
  https://groups.google.com/d/forum/stan-users



Motivation for Stan

Stan 主要用于 Bayesian 统计分析

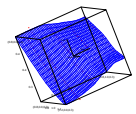
- Fit rich Bayesian statistical models
- The Process
  - Create a statistical model
  - Perform inference on the model
  - Evaluate
- Difficulty with models of interest in existing tools



Motivation (cont.)

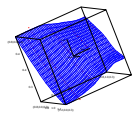
Stan 的优点在于

- Mainly used for multilevel generalized linear models
- Using C++ to overcome the cost of interpretation
- Better sampler using for highly correlated in the posterior
- Hamiltonian Monte Carlo(HMC) sampling



Bayesian 统计基本知识

- 贝叶斯统计的基本思想是，任何一个未知量 θ 都是一个随机变量，关于该随机变量的信息可以来自两个方面
- 样本的信息 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$
- 先验知识的信息 $\pi(\theta)$
- 得到关于 θ 的后验分布 $\pi(\theta|x)$



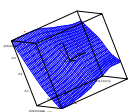
Bayesian 统计基本知识

假设 θ 是未知参数，在给定样本 $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ 的条件下，关于参数 θ 的贝叶斯推断可以归纳如下：

- 根据已有的先验信息，确定 θ 的先验分布 $\pi(\theta)$
- 在给定样本 $\mathbf{x}$ 的条件下，θ 的后验分布为

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta, \mathbf{x})}{m(x)} = \frac{f(\mathbf{x}|\theta) \cdot \pi(\theta)}{\int f(\mathbf{x}|\theta) \cdot \pi(\theta) \mathrm{d}\theta}$$

其中，密度函数 $f(\mathbf{x}|\theta)$ 表示在 θ 给定的条件下，$\mathbf{x}$ 的条件分布。



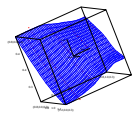
先验分布的选择

- 无信息先验分布：

  无信息先验分布是指除参数 θ 的取值范围和 θ 在总体分布中的地位之外，再也不包含 θ 的任何信息的先验分布。

- 当对参数 θ 一无所知时，可以认为 θ 的取值均匀地分布在其变化范围内，即

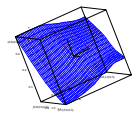
$$\pi(\theta) = \begin{cases} c, & \theta \in \Theta \\ 0, & \theta \notin \Theta \end{cases}$$



贝叶斯估计方法

马尔科夫链蒙特卡洛（MCMC）随机模拟

- Gibbs 抽样
- Metropolis-Hastings 抽样
- Hamiltonian Monte Carlo



模型诊断

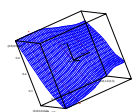
- 轨迹图、自相关图
- 有效样本量(Effective sample size)

$$\text{ESS} = \frac{\text{T}}{1 + 2\sum_{k=0} \rho_{(k)}(\theta)}$$

- 潜在尺度缩减因子

$$\hat{R} = \sqrt{\frac{\frac{N-1}{N}\text{W} + \frac{1}{N}B}{\text{W}}}$$

其中，W 和 B 分别为样本 $\theta_m^{(n)}$ 的组内和组间方差。潜在尺度缩减因子 $\hat{R}$ 是马尔科夫链的组内平均方差相对于组间方差的比率



模型评价

WAIC统计量

- lpd: log pointwise predictive density

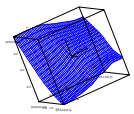
$$lpd = \sum_{i=1}^{n} \log p(y_i|y) = \sum_{i=1}^{n} \int p(y_i|\theta)p(\theta|y)d\theta$$

$$\hat{lpd} = \sum_{i=1}^{n} \log \left[\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^S) \right]$$

- p_{waic}: effective numbers of parameters

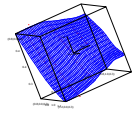
$$p_{waic} = \sum_{i}^{n} \mathsf{Var}_{post} \left[\log p(y_i|\theta) \right]$$

$$\hat{p}_{waic} = \sum_{i=1}^{n} \mathsf{V}_{S=1}^{S} \left[\log p(y_i|\theta^S) \right]$$



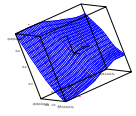
The Stan Language

- Program Blocks
  - - Data
  - - Transformed data
  - - Parameters
  - - Transformed parameters
  - - Model
  - - Generated quantities



The Stan Language

- Data Types
  - 连续性变量：real
  - 整数：int
  - 列，行: vector, row vector
  - 矩阵 $[n, k]$ Matrix
  - 数组 Arrays
  - 施加限制，如 <lower=0, upper=1>
- Vector 与 Matrix 只能用于 real　下述条件只能使用
  - Matrix arithmetic operations (e.g., matrix multiplication)
  - Linear algebra functions (e.g., eigenvalues and determinants)
  - Multivariate function parameters and outcomes (e.g., multivariate normal distribution arguments)

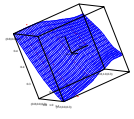


The Stan Language

Name	Function
`print`	print the summary for parameters obtained using all chains
`summary`	summarize the sample from all chains and individual chains for parameters
`plot`	plot the inferences (intervals, medians, split $\hat{R}$) for parameters
`traceplot`	plot the traces of chains
`extract`	extract samples of parameters
`get_stancode`	extract the model code in Stan modeling language
`get_stanmodel`	extract the `stanmodel` object
`get_seed`	get the seed used for sampling
`get_inits`	get the initial values used for sampling
`get_posterior_mean`	get the posterior mean for all parameters
`get_logposterior`	get the log posterior (that is, `lp__`)
`get_sampler_params`	get parameters used by the sampler such as `treedepth` of NUTS
`get_adaptation_info`	get adaptation information of the sampler
`get_num_upars`	get the number of parameters on unconstrained space
`unconstrain_pars`	transform parameter to unconstrained space
`constrain_pars`	transform parameter from unconstrained space to its defined space
`log_prob`	evaluate the log posterior for parameter on unconstrained space
`grad_log_prob`	evaluate the gradient of the log posterior for parameter on unconstrained space

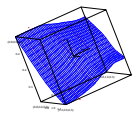
`as.array`
`as.matrix`
`as.data.frame`
	extract the samples excluding warmup to a three dimension array, matrix, data.frame

`pairs`	make a matrix of scatter plots for the samples of parameters



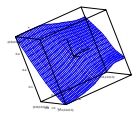
2. Fitting Distribution in R

[例8-3]: 假设伽马分布的均值为 15，形状参数（shape）和比率参数（rate）分别为 30 和 2 。请模拟 1000 个服从该伽马分布的随机数，并基于这组随机数应用 Rstan 估计伽马分布的两个参数。



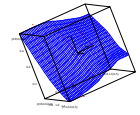
3. Linear Regression Models

[例8-4]：假设 y 服从正态分布，标准差为 2，均值可以表示为三个协变量的线性函数，即 $\mu = 10 + 0.2 \times x_1 - 0.3 \times x_2 + 0.4 \times x_3$，其中三个协变量都服从标准正态分布。请模拟 1000 个因变量和协变量的观察值，并基于这组模拟数据应用 Rstan 建立线性回归模型。



Generalized Linear Models

- 广义线性模型是线性模型的扩展，由正态分布扩展到指数分布族
- 拟合偏态分布，如伽马分布、逆高斯分布等
- 拟合离散分布，如泊松分布、负二项分布等
- 在 R 软件中运用的包主要有： glm, gamlss 等



Generalized Linear Models

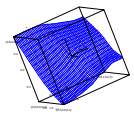
假设随机变量 Y 服从指数分布簇:

$$f_{Y_i}(y|\eta_i,\phi)=\exp\left[\frac{y\eta_i-b(\eta_i)}{\phi}+c(y,\phi)\right]$$

其中 ϕ 为离散参数, η_i 为自然参数
期望和方差分别表示为:

$$E(Y_i)=b'(\eta_i)$$
$$\mathrm{Var}(Y_i)=V(\mu_i)\phi$$

上式中，ϕ 为常数，$V(\cdot)$ 表示方差函数。



伽马回归模型

假设 Y 服从伽马分布, 即

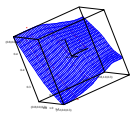
$$f(y|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp(\beta y)$$

期望与方差为

$$\mu = \frac{\alpha}{\beta}$$
$$\mathrm{Var}(Y) = \frac{\alpha}{\beta^2}$$

连接函数与协变量:

$$\log(\mu) = \mathbf{X}\boldsymbol{\beta}$$



Gamma Regression in R

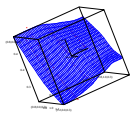
模拟数据:连续因变量（伽马分布）

数据包含 1000 个观测值，记为 y_i. 只包含一个协变量 X_1。其中 y_i 服从伽马分布，期望为 μ_i, 即：

$$y_i \sim \text{Gamma}(shape, rate)$$
$$y_i \sim \text{Gamma}(\mu, \sigma)$$

均值与协变量的关系为：

$$\mu_i = \exp(5 + 1.2 \times x_1 + 1.3 \times x_2 + 1.4 \times x_3)$$
$$x_1, x_2, x_3 \sim \text{exponential}(0.1)$$

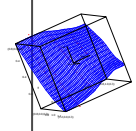


流量三角形

- 上三角有 171 个已知的观测值，表示已付增量赔款，数据来自Dong（2015）

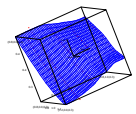
表 2 增量赔款数据

事故年	进展年																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3323	8332	9572	10172	7631	3855	3252	4433	2188	333	199	692	311	0.01	405	293	76	14
2	3785	10342	8330	7849	2839	3577	1404	1721	1065	156	35	259	250	420	6	1	0.01	
3	4677	9989	8746	10228	8572	5787	3855	1445	1612	626	1172	589	438	473	370	31		
4	5288	8089	12839	11829	7560	6383	4118	3016	1575	1985	2645	266	38	45	115			
5	2294	9869	10242	13808	8775	5419	2424	1597	4149	1296	917	295	428	359				
6	3600	7514	8247	9327	8584	4245	4096	3216	2014	593	1188	691	368					
7	3642	7394	9838	9733	6377	4884	11920	4188	4492	1760	944	921						
8	2463	5033	6980	7722	6702	7834	5579	3622	1300	3069	1370							
9	2267	5959	6175	7051	8102	6339	6978	4396	3107	903								
10	2009	3700	5298	6885	6477	7570	5855	5751	3871									
11	1860	5282	3640	7538	5157	5766	6862	2572										
12	2331	3517	5310	6066	10149	9265	5262											
13	2314	4487	4112	7000	11163	10057												
14	2607	3952	8228	7895	9317													
15	2595	5403	6579	15546														
16	3155	4974	7961															
17	2626	5704																
18	2827																	



模型拟合

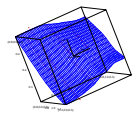
- 事故年和进展年作为解释变量
- 假设增量赔款服从伽马分布
- 分别用极大似然估计方法和贝叶斯方法得到模型的参数
- 对下三角进行预测



总结

在运用 rstan 中需要注意的几点

- 在贝叶斯分析框架中，先验分布的选择很重要。当模型参数不收敛时，可以根据研究经验选择合适的弱信息或者信息先验分布
- Rstan 中如果不事先设定先验分布，程序默认为无信息的先验分布
- 当模型通过诊断时，才能根据抽取的样本进行统计推断和预测



Thank you !