厦門大學 数学科学学院
SCHOOL OF MATHEMATICAL SCIENCES XIAMEN UNIVERSITY

统计计算小组作业汇报

GROUP 4

小组成员：何啻 陈飞亦 王琳玥 谢奕童 刘诗琪

QUESTION 3

0-1膨胀负二项分布模型参数估计

模型分析

▼ **模型介绍：**

0-1膨胀负二项回归模型（Zero-and-One-Inflated Negative Binomial Regression Model）通常用于处理计数变量中存在过多的0值的情况，这些0值可能是"**真实的零**"（实际计数为0）和"**伪零**"（由于其他因素导致计数未发生），为了拟合这种情形的数据，于是引入**伯努利与负二项分布**混合的ZOINB模型： $Y_i = (1 - D_i)M_i + D_iZ_i$ 。

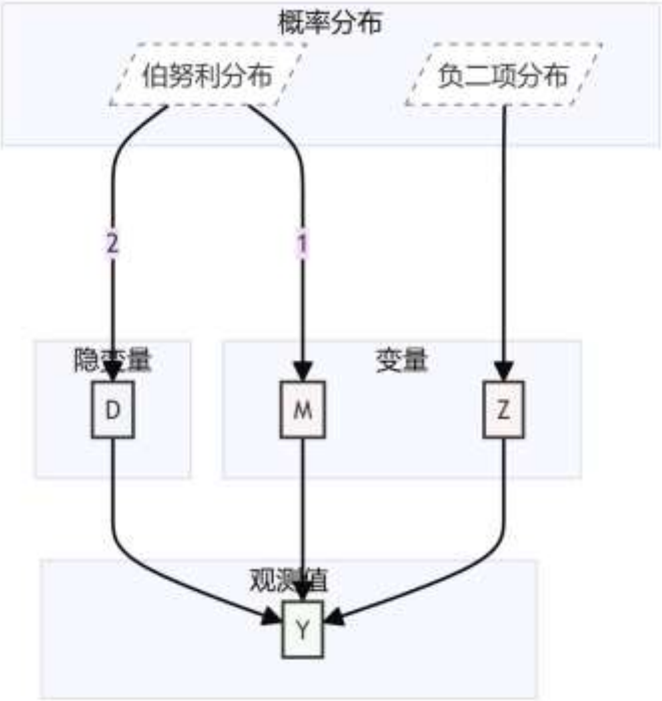
在该模型中，观测值 Y_i 由两个**潜在随机变量** M_i 和 Z_i 通过隐变量 D_i 决定。

其中 D_i 服从**伯努利分布**，取值为 0 或 1。

- 当 $D_i = 1$ 时， $Y_i = M_i$ ，其中 M_i 服从**伯努利分布**；
- 当 $D_i = 0$ 时， $Y_i = Z_i$ ，其中 Z_i 服从**负二项分布。**

因此，观测序列（$Y_1, \cdots, Y_n$）实际上是由潜在序列（$M_1, \cdots, M_n$）和（$Z_1, \cdots, Z_n$）根据隐变量序列（$D_1, \cdots, D_n$）选择组合而成。

每个观测值 Y_i 都是其对应位置的 M_i 或 Z_i 之一，具体取决于隐变量 D_i 的取值。



模型：

- **观测值**：$Y_i = (1 - D_i)M_i + D_iZ_i$
- **混合成分1**（伯努利分布）：$M_i \sim B(1, p)$
- **混合成分2**（负二项分布）：$Z_i \sim NB(\mu_i, \phi)$
- **混合来源标签**：$D_i \sim B(1, 1 - \pi_0)$，为了和先验概率符号 $\pi(\cdot)$ 区分，后文均将题目中的 π 转换为 π_0

便捷起见，下记全参数为 θ

数据：

- 观测数据：$Y = (Y_1, \cdots, Y_n)$
- 隐含数据：
  - $D = (D_1, \cdots, D_n)$，且 $Y_i > 1$ 时 $D_i = 1$
  - $M = (M_1, \cdots, M_n)$
  - $Z = (Z_1, \cdots, Z_n)$
- 完全数据：(Y, D, M, Z)

似然函数

- **似然函数：**
  - **完全数据的单样本概率函数**就可以写为：

$$p(Y_i, D_i, M_i, Z_i|\theta) = p(Y_i, M_i, Z_i|D_i, \theta)p(D_i)$$

  - **完全数据的似然函数**：

$$L(Y, D, M, Z|\theta) = \prod_{i=1}^{n} p(Y_i, D_i, M_i, Z_i|\theta) = \prod_{i=1}^{n} p(Y_i|D_i, M_i, Z_i, \theta)p(D_i) \tag{1}$$

先验分布

先验分布：

- 参数向量 $\tilde{\beta}=(\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\beta}_2)^\top$ ：多维正态分布 $\tilde{\beta}\sim N_3(\mu_{\tilde{\beta}},\sigma_{\tilde{\beta}}^2 I_3)$ ，其中 $\mu_{\tilde{\beta}}$ 是已知向量，$\sigma_{\tilde{\beta}}^2$ 是已知常数。
- 参数 ϕ ：Gamma分布 $\phi\sim Gamma(e_0,f_0)$ ，其中 e_0 是形状参数，f_0 是速率参数
- 参数 π_0 ：区间 $[0,1]$ 上的均匀分布，即 $\pi_0\sim U(0,1)$ 。
- 参数 p ：$[0,1]$ 上的均匀分布，即 $p\sim U(0,1)$ 。

进一步，我们假设 $\tilde{\beta},\phi,\pi_0$ 和 p 相互独立。

故 $(\tilde{\beta},\phi,\pi_0,p)$ 的联合先验分布 $\pi(\tilde{\beta},\phi,\pi_0,p)=\pi(\tilde{\beta})\pi(\phi)\pi(\pi_0)\pi(p)$

后验分布

得到数据扩充下 $(\tilde{\beta}, \phi, \pi_0, p)$ 的**后验分布正比形式为**:

$$
\begin{aligned}
\pi((\tilde{\beta}, \phi, \pi_0, p)|Y, D, M, Z) \propto & \prod_{i=1}^{n} \left[\frac{\Gamma(\phi + Z_i)}{\Gamma(\phi)}\right]^{D_i} \times \prod_{i=1}^{n} \frac{\{\exp(X_i^T \tilde{\beta})\}^{Z_i D_i}}{\{1 + \exp(X_i^T \tilde{\beta})\}^{(Z_i+\phi)D_i}} \\
& \times \left[p^{\sum_{i=1}^{n} M_i(1-D_i)} (1-p)^{\sum_{i=1}^{n}(1-M_i)(1-D_i)}\right] \\
& \times (1-\pi_0)^{\sum_{i=1}^{n} D_i} {\pi_0}^{n-\sum_{i=1}^{n} D_i} \\
& \times \pi(\tilde{\beta})\pi(\phi)\pi(\pi_0)\pi(p)
\end{aligned} \tag{6}
$$

满条件分布——六个参数

(式中"-"表示其余给定参数)

参数	正比形式
$\tilde{\beta}$	$\pi(\tilde{\beta}\|-,Y,D,M,Z) \propto \prod_{i=1}^{n} \frac{\{\exp(X_i^T\tilde{\beta})\}^{Z_iD_i}}{\{1+\exp(X_i^T\tilde{\beta})\}^{(Z_i+\phi)D_i}} \times \pi(\tilde{\beta}) \quad (11)$
ϕ	$\pi(\phi\|-,Y,D,M,Z) \propto \prod_{i=1}^{n} \left[\frac{\Gamma(\phi+Z_i)}{\Gamma(\phi)}\right]^{D_i} \times \prod_{i=1}^{n} \{1+\exp(X_i^T\tilde{\beta})\}^{-\phi D_i} \times \pi(\phi) \quad (12)$
π_0	$\pi(\pi_0\|-,Y,D) \propto (1-\pi_0)^{\sum_{i=1}^{n} D_i} {\pi_0}^{n-\sum_{i=1}^{n} D_i} \times \pi(\pi_0) \quad (13)$
p	$\pi(p\|-,Y,D) \propto \left[p^{\sum_{i=1}^{n} M_i(1-D_i)}(1-p)^{\sum_{i=1}^{n}(1-M_i)(1-D_i)}\right] \times \pi(p) \quad (14)$

满条件分布——隐含数据D, M, Z

观测数据：
$Y_i = (1 - D_i)M_i + D_iZ_i$

- **混合成分1**：$M_i \sim B(1, p)$
- **混合成分2**：$Z_i \sim NB(\mu_i, \phi)$
- **混合来源标签**：$D_i \sim B(1, 1 - \pi_0)$

均需要在后续Gibbs抽样中迭代更新

Y_i 观测到不同数值时，背后的事件组合：

	D_i	M_i	Y_i	发生该情况的概率
情况1:	0	0	$Y_i = (1 - 0)0 + 0Z_i = 0$	$\pi_0(1 - p)$
情况2:	0	1	$Y_i = (1 - 0)1 + 0Z_i = 1$	$\pi_0 p$
情况3:	1	0	$Y_i = (1 - 1)0 + 1Z_i = Z_i$	$(1 - \pi_0)(1 - p)$
情况4:	1	1	$Y_i = (1 - 1)1 + 1Z_i = Z_i$	$(1 - \pi_0)p$

$$\begin{cases} \{Y_i = 0\} \Leftrightarrow \{D_i = 0, M_i = 0\} \bigcup \{D_i = 1, Z_i = 0\} \\ \{Y_i = 1\} \Leftrightarrow \{D_i = 0, M_i = 1\} \bigcup \{D_i = 1, Z_i = 1\} \\ \{Y_i = k\} \Leftrightarrow \{D_i = 1, Z_i = k\}(k > 1) \end{cases}$$

满条件分布——隐含数据D, M, Z

隐含数据 D, M, Z 的概率公式 $\{Y_i = 0\} \Leftrightarrow \{D_i = 0, M_i = 0\} \bigcup \{D_i = 1, Z_i = 0\}$

$$P(D_i = a, M_i = b, Z_i = k | Y_i = 0, \theta) = \begin{cases} \frac{(1-\pi_0)pP(Z_i=0)}{\pi_0(1-p)+(1-\pi_0)P(Z_i=0)}, & \text{if } k = 0, a = b = 1, \\ \frac{(1-\pi_0)(1-p)P(Z_i=0)}{\pi_0(1-p)+(1-\pi_0)P(Z_i=0)}, & \text{if } k = 0, a = 1, b = 0, \\ \frac{\pi_0(1-p)P(Z_i=k)}{\pi_0(1-p)+(1-\pi_0)P(Z_i=0)}, & \text{if } a = b = 0, k = 0, 1, \cdots, \\ 0, & \text{otherwise}; \end{cases}$$

式（11）-1

满条件分布——隐含数据D, M, Z

隐含数据 D, M, Z 的概率公式 $\{Y_i = 1\} \Leftrightarrow \{D_i = 0, M_i = 1\} \bigcup \{D_i = 1, Z_i = 1\}$

$$P(D_i = a, M_i = b, Z_i = k \mid Y_i = 1, \theta) = \begin{cases} \frac{(1-\pi_0)pP(Z_i=1)}{\pi_0 p + (1-\pi_0)P(Z_i=1)}, & \text{if } k = a = b = 1, \\ \frac{(1-\pi_0)(1-p)P(Z_i=1)}{\pi_0 p + (1-\pi_0)P(Z_i=1)}, & \text{if } k = a = 1, b = 0, \\ \frac{\pi_0 pP(Z_i=k)}{\pi_0 p + (1-\pi_0)P(Z_i=1)}, & \text{if } a = 0, b = 1, k = 0, 1, \cdots, \\ 0, & \text{otherwise}; \end{cases}$$

式（11）- 2

满条件分布——隐含数据D, M, Z

隐含数据 D, M, Z 的概率公式

$$\{Y_i = k\} \Leftrightarrow \{D_i = 1, Z_i = k\}(k > 1)$$

⬇

$$P(D_i = a, M_i = b, Z_i = k|Y_i = k, \theta) = \begin{cases} 1-p, & \text{if } a = b = 0, k = 2, 3, \cdots, \\ p, & \text{if } a = 0, b = 1, k = 2, 3, \cdots, \\ 0, & \text{otherwise.} \end{cases}$$

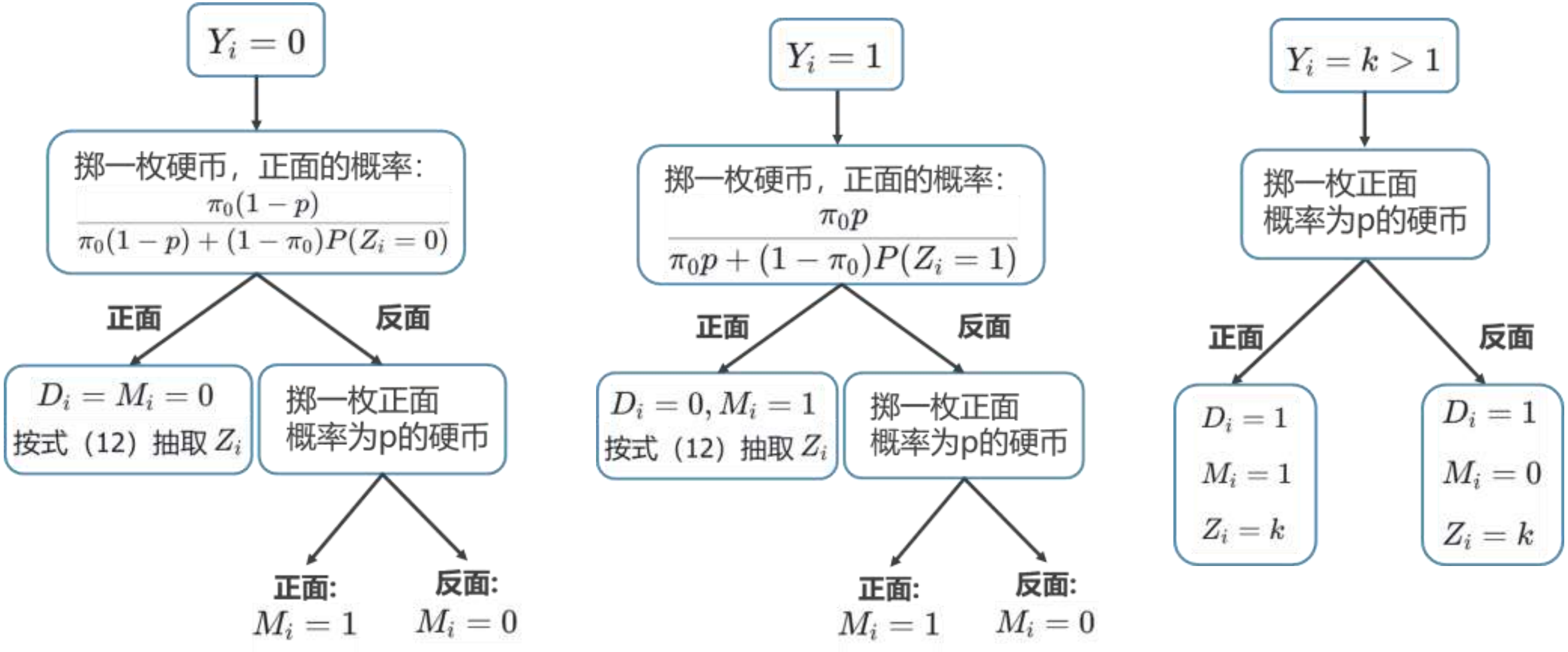
式（11）- 3

其中 $P(Z_i = k|\theta)$ 由以下概率分布列生成:

$$P(Z_i = k|\mu_i, \phi) = \frac{\Gamma(\phi + k)}{\Gamma(k+1)\Gamma(\phi)} \frac{\{\exp(X_i^T\tilde{\beta})\}^k}{\{1 + \exp(X_i^T\tilde{\beta})\}^{(k+\phi)}}, k = 0, 1, 2, \cdots \qquad (12)$$

抽样方法01——隐含数据D, M, Z

D_i, M_i, Z_i的更新式可以根据式（11）进行判定情形，然后**利用离散分布**进行抽样



抽样方法02——参数π_0, p

由满条件分布式:

$$
\begin{aligned}
\pi(\pi_0|-,Y,D) &\propto (1-\pi_0)^{\sum_{i=1}^n D_i} {\pi_0}^{n-\sum_{i=1}^n D_i} \times \pi(\pi_0) \\
&\sim \mathrm{Beta}(n+1-\sum_{i=1}^n D_i, \sum_{i=1}^n D_i + 1)
\end{aligned} \tag{9}
$$

$$
\begin{aligned}
\pi(p|-,Y,D) &\propto \left[p^{\sum_{i=1}^n M_i(1-D_i)}(1-p)^{\sum_{i=1}^n (1-M_i)(1-D_i)}\right] \times \pi(p) \\
&\sim \mathrm{Beta}(\sum_{i=1}^n M_i(1-D_i)+1, \sum_{i=1}^n (1-M_i)(1-D_i)+1)
\end{aligned} \tag{10}
$$

可以直接利用**Beta分布**对参数π_0, p进行抽样

抽样方法03——参数$\tilde{\beta}$

3 式（7）（8）并不是常见的分布，从中进行直接抽样比较困难。
但由于Gibbs对其他参数更新的便捷性，我们在抽样时仍考虑Gibbs抽样。

我们在对参数 $\tilde{\beta},\phi$ 更新时考虑采用**M-H抽样算法**。根据文献[6][7]，可知在Gibbs抽样中混合**对某部分参数M-H抽样的可行性**。

$\tilde{\beta}$

Article

Bayesian Analysis Using Joint Progressive Type-II Censoring Scheme

Mohamed G. M. Ghazal [1,2], Mustafa M. Hasaballah [3,*], Rashad M. EL-Sagheer [4,5], Oluwafemi Samson Balogun [6] and Mahmoud E. Bakr [7]

1 Department of Mathematics, Faculty of Science, Minia University, Minia 61519, Egypt
2 Department of Mathematics, College of Education, University of Technology and Applied Sciences, Al-Rustaq 329, Oman
3 Marg Higher Institute of Engineering and Modern Technology, Cairo 11721, Egypt
4 Mathematics Department, Faculty of Science, Al-Azhar University, Cairo 11884, Egypt
5 High Institute of Computer and Management Information System, First Statement, New Cairo 11865, Egypt
6 Department of Computing, University of Eastern Finland, FI-70211 Kuopio, Finland
7 Department of Statistics and Operations Research, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia
* Correspondence: mustafamath7@yahoo.com

文献[6]

ϕ

Article

Parameter and Reliability Inferences of Inverted Exponentiated Half-Logistic Distribution under the Progressive First-Failure Censoring

Fengshi Zhang and Wenhao Gui *

Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China; 17271109@bjtu.edu.cn
* Correspondence: whgui@bjtu.edu.cn

check for updates

Received: 13 April 2020; Accepted: 27 April 2020; Published: 3 May 2020

Abstract: Using progressive first-failure censored samples, we mainly study the inferences of the unknown parameters and the reliability and failure functions of the Inverted Exponentiated Half-Logistic distribution. The progressive first-failure censoring is an extension and improvement of progressive censoring, which is of great significance in the field of lifetime research. Besides maximum likelihood

文献[7]

抽样方法03——参数$\tilde{\beta}$

参数	正比形式
$\tilde{\beta}$	$\pi(\tilde{\beta}\mid -, Y, D, M, Z) \propto \prod_{i=1}^{n} \frac{\{\exp(X_i^T\tilde{\beta})\}^{Z_i D_i}}{\{1+\exp(X_i^T\tilde{\beta})\}^{(Z_i+\phi)D_i}} \times \pi(\tilde{\beta}) \qquad (11)$

式（11）并不是常见的分布，从中进行直接抽样比较困难，在更新该维度的参数时考虑方法：

1. 使用Metropolis - Hastings算法
2. 引入 Pólya–Gamma 潜变量

抽样方法03——参数$\tilde{\beta}$

1. M-H抽样

可行性分析:

根据文献[6][7]，可知在Gibbs抽样中混合**对某部分参数M-H抽样的可行性**。

Article

Bayesian Analysis Using Joint Progressive Type-II Censoring Scheme

Mohamed G. M. Ghazal [1,2], Mustafa M. Hasaballah [3,*], Rashad M. EL-Sagheer [4,5], Oluwafemi Samson Balogun [6] and Mahmoud E. Bakr [7]

1 Department of Mathematics, Faculty of Science, Minia University, Minia 61519, Egypt
2 Department of Mathematics, College of Education, University of Technology and Applied Sciences, Al-Rustaq 329, Oman
3 Marg Higher Institute of Engineering and Modern Technology, Cairo 11721, Egypt
4 Mathematics Department, Faculty of Science, Al-Azhar University, Cairo 11884, Egypt
5 High Institute of Computer and Management Information System, First Statement, New Cairo 11865, Egypt
6 Department of Computing, University of Eastern Finland, FI-70211 Kuopio, Finland
7 Department of Statistics and Operations Research, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia
* Correspondence: mustafamath7@yahoo.com

文献[6]

Article

Parameter and Reliability Inferences of Inverted Exponentiated Half-Logistic Distribution under the Progressive First-Failure Censoring

Fengshi Zhang and Wenhao Gui *

Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China; 17271109@bjtu.edu.cn
* Correspondence: whgui@bjtu.edu.cn

Received: 13 April 2020; Accepted: 27 April 2020; Published: 3 May 2020

check for updates

Abstract: Using progressive first-failure censored samples, we mainly study the inferences of the unknown parameters and the reliability and failure functions of the Inverted Exponentiated Half-Logistic

文献[7]

抽样方法03——参数$\tilde{\beta}$

1. M-H抽样

$$p(x,x') = \boxed{q(x,x')}\,\boxed{\alpha(x,x')}$$

建议分布：

为了**简化模型**

取潜在的转移核 $q(x,x')$ 为 $N(x,\sigma_0^2)$

其中 σ_0^2 为自行指定的超参数，设置为 0.01

接受概率：

为了使目标 $\pi(x)$ 成为平稳分布

选择 $\alpha(\cdot,\cdot)$：

$$\alpha(x,x') = \min\left(1, \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}\right)$$

$$X^{(t+1)} = \begin{cases} x', & u \leqslant \alpha(x,x') \quad \text{接受转移} \\ x = X^{(t)}, & u > \alpha(x,x') \quad \text{拒绝转移，停留在原地} \end{cases}$$

$$p(x,x') = \begin{cases} q(x,x'), & \pi(x')q(x',x) \geqslant \pi(x)q(x,x') \\ q(x',x)\frac{\pi(x')}{\pi(x)}, & \pi(x')q(x',x) < \pi(x)q(x,x') \end{cases}$$

抽样方法03——参数$\tilde{\beta}$

1. M-H抽样

在每一次发生转移/迭代时：

step1： 链在时刻 t 处于状态 x ，即 $X^{(t)} = x$ 。

step2： 由 $q(x,\cdot)$ 产生一个潜在的转移 $x \rightarrow x'$

step3： 然后根据概率 $\alpha(x,x')$ 决定是否转移。

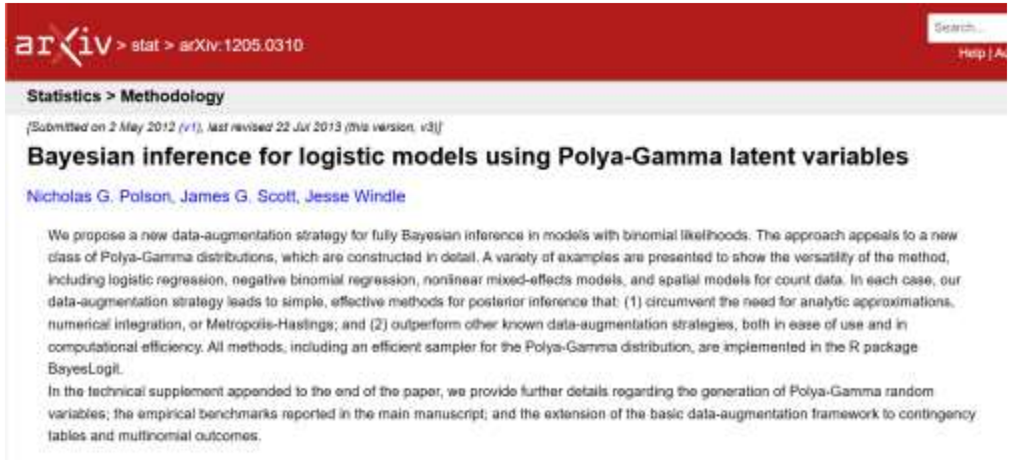
抽样方法03——参数$\tilde{\beta}$

2. 引入 Pólya−Gamma 潜变量

M-H拒绝率高，抽样效率低 → 引入 Pólya−Gamma 潜变量

引理[5]：设 $p(\omega)$ 为Pólya - Gamma分布 $PG(b,0)$ （ $b>0$ ）的概率密度函数，对于任意实数 $a \in R$ ，有 $\frac{\{e^{\psi}\}^a}{\{1+e^{\psi}\}^b} = 2^{-b}e^{\kappa\psi}\int_0^{\infty} e^{-\frac{\omega\psi^2}{2}}p(\omega)d\omega$ ，其中 $\kappa = a - \frac{b}{2}$ 。

根据以上引理，我们引入 Pólya−Gamma 变量，结合 Pólya−Gamma 分布导出的**条件高斯分布**对 $\tilde{\beta}$ 进行抽样，得到高效率的后验样本。


arXiv > stat > arXiv:1205.0310

Search... Help | A

Statistics > Methodology

[Submitted on 2 May 2012 (v1), last revised 22 Jul 2013 (this version, v3)]

Bayesian inference for logistic models using Polya-Gamma latent variables

Nicholas G. Polson, James G. Scott, Jesse Windle

We propose a new data-augmentation strategy for fully Bayesian inference in models with binomial likelihoods. The approach appeals to a new class of Polya-Gamma distributions, which are constructed in detail. A variety of examples are presented to show the versatility of the method, including logistic regression, negative binomial regression, nonlinear mixed-effects models, and spatial models for count data. In each case, our data-augmentation strategy leads to simple, effective methods for posterior inference that: (1) circumvent the need for analytic approximations, numerical integration, or Metropolis-Hastings; and (2) outperform other known data-augmentation strategies, both in ease of use and in computational efficiency. All methods, including an efficient sampler for the Polya-Gamma distribution, are implemented in the R package BayesLogit.

In the technical supplement appended to the end of the paper, we provide further details regarding the generation of Polya-Gamma random variables; the empirical benchmarks reported in the main manuscript; and the extension of the basic data-augmentation framework to contingency tables and multinomial outcomes.

抽样方法03——参数$\tilde{\beta}$

2. 引入 Pólya-Gamma 潜变量

引理[5]：设 $p(\omega)$ 为Pólya - Gamma分布 $PG(b,0)$ （ $b>0$ ）的概率密度函数，对于任意实数 $a\in R$ ，有 $\frac{\{e^{\psi}\}^a}{\{1+e^{\psi}\}^b}=2^{-b}e^{\kappa\psi}\int_0^{\infty}e^{-\frac{\omega\psi^2}{2}}p(\omega)d\omega$ ，其中 $\kappa=a-\frac{b}{2}$ 。

$$\pi(\tilde{\beta}|-,Y,D,M,Z)$$
$$\propto\prod_{i=1}^{n}\frac{\{\exp(X_i^T\tilde{\beta})\}^{Z_iD_i}}{\{1+\exp(X_i^T\tilde{\beta})\}^{(Z_i+\phi)D_i}}\times\pi(\tilde{\beta})$$

$$\pi(\tilde{\beta}|-,Y,D,M,Z)\propto\pi(\tilde{\beta})\prod_{i=1}^{n}\frac{\{exp(X_i^T\tilde{\beta})\}^{Z_iD_i}}{\{1+exp(X_i^T\tilde{\beta})\}^{(Z_i+\phi)D_i}}$$
$$\propto\pi(\tilde{\beta})\times\prod_{i=1}^{n}exp(\tilde{\kappa}_iX_i^T\tilde{\beta})\times\int_0^{\infty}exp(-\frac{\omega_i(X_i^T\tilde{\beta})^2}{2})p(\omega_i)d\omega_i$$

其中： $\tilde{\kappa}_i=\frac{(Z_i-\phi)D_i}{2}$ ， ω_i 服从 $PG((Z_i+\phi)D_i,0)$ 分布。

2. 引入 Pólya-Gamma 潜变量

记 $\omega=(\omega_1,\omega_2,\cdots,\omega_n)$ 为Pólya - Gamma变量。

若 ω_i 已经从 $PG((Z_i+\phi)D_i,0)$ 分布中抽样得到，对于给定的 ω ，有：

$$
\begin{aligned}
\underbrace{\boxed{\pi(\tilde{\beta}|-,Y,D,M,Z,\omega)}}_{\text{条件高斯分布}} &\propto \pi(\tilde{\beta})\times\prod_{i=1}^{n} exp\left\{\tilde{\kappa}_i X_i^T\tilde{\beta}-\frac{\omega_i(X_i^T\tilde{\beta})^2}{2}\right\} \\
&\propto \pi(\tilde{\beta})\times\prod_{i=1}^{n} exp\left\{-\frac{\omega_i}{2}(\frac{\tilde{\kappa}_i}{\omega_i}-X_i^T\tilde{\beta})^2\right\} \\
&\propto \pi(\tilde{\beta})\times exp\left\{-\frac{1}{2}(\lambda-X\tilde{\beta})^T\Omega(\lambda-X\tilde{\beta})\right\}
\end{aligned} \tag{13}
$$

其中： $\lambda=(\frac{\tilde{\kappa}_1}{\omega_1},\frac{\tilde{\kappa}_2}{\omega_2},\cdots,\frac{\tilde{\kappa}_n}{\omega_n})$ ， $\Omega=diag(\omega_1,\omega_2,\cdots,\omega_n)$ ，
$X_{n\times 3}=(X_1^T,X_2^T,\cdots,X_n^T)^T$ 。

抽样方法03——参数$\tilde{\beta}$

2. 引入 Pólya-Gamma 潜变量

因此，由式 (13) 抽样得到 $\tilde{\beta}$ 的后验样本，公式为:

$$\omega_i \sim PG((Z_i+\phi)D_i, 0), i=1,2,\cdots,n$$

$$\tilde{\beta}|-, Y, D, M, Z, \omega \sim N\left(\tilde{M}_\omega, \tilde{H}_\omega\right) \tag{14}$$

其中 $\tilde{H}_\omega = \left(X^T \Omega X + \sigma_{\tilde{\beta}}^{-2} I_s^{-1}\right)^{-1}$，$\tilde{M}_\omega = \tilde{H}_\omega \left(X^T \tilde{\kappa} + \sigma_{\tilde{\beta}}^{-2} I_s^{-1} \mu_{\tilde{\beta}}\right)$，

$\tilde{\kappa} = (\frac{(Z_1-\phi)D_1}{2}, \frac{(Z_2-\phi)D_2}{2}, \cdots, \frac{(Z_n-\phi)D_n}{2})$。

由此即可实现对 $\tilde{\beta}$ 的抽样。

注：$D_i = 0$时需要特殊处理

抽样方法04——参数∅

对参数 ϕ 的更新
1. 使用Metropolis - Hastings算法
2. 使用两阶段Gibbs采样

> 一个经典事实是负二项分布等价于伽马 - 泊松混合分布：
>
> 我们可以通过先抽取 $\lambda \sim Gamma(\phi, (1-p)/p)$ ，然后生成 $y|\lambda \sim Pois(\lambda)$ 来得到 $y|\phi, p \sim NB(\phi, p)$ 。负二项分布也可以在复合泊松表示下进行扩充，这样通过给参数 ϕ 和 p 赋予合适的共轭先验，就有可能以一种易于处理的方式抽取它们的后验分布，见Zhou和Car in（2015）[9]。
>
> 具体来说，给定负二项模型 $y_j|\phi, p \overset{iid}{\sim} NB(\phi, p)$ ，对于 $j = 1, \cdots, n$ ，先验设定为 $p \sim Beta(a_0, b_0)$ ， $\phi \sim Gamma(e_0, f_0)$ ，其中 e_0 和 f_0 是形状和速率参数。然后通过迭代应用以下吉布斯采样步骤来获得 p 和 ϕ 的后验分布：

(注：这里符号 ϕ, μ 均与上述一致，此处的 p 为上述等价更换形式后的符号 γ)

抽样方法04——参数∅

在**零膨胀负二项回归模型的Gibbs抽样过程**中，对参数 ϕ **更新规则**如下：

首先根据 $D^{(t-1)}$ 筛选出 $D_i^{(t-1)}=1$ 对应的 Y_i，记这些被判定来源于负二项分布的 Z_i 的下标为 $n_j, j=1,2,\cdots,h$。

然后利用这些数据进行更新：

😸

$$
\begin{aligned}
&\textbf{step1:}\ \ (l_j|-,Y,D^{(t-1)},M^{(t-1)},Z^{(t-1)}) \sim CRT(Y_{n_j},\phi^{(t-1)}) \\
&\textbf{step2:}\ \ \gamma_{n_j}^{(t-1)} = \frac{\mu_{n_j}^{(t-1)}}{\mu_{n_j}^{(t-1)}+\phi^{(t-1)}} \\
&\textbf{step3:}\ \ (\phi^{(t+1)}|-,Y,D^{(t-1)},M^{(t-1)},Z^{(t-1)})) \sim Gamma(e_0+\sum_j^h l_j, f_0-\sum_{j=1}^h \log(1-\gamma_{n_j}^{(t-1)}))
\end{aligned}
\tag{13}
$$

在迭代结束后，可以考虑将估计参数 $\hat{\phi}$ 进行取整操作。

抽样方法04——参数∅

step1： $(l_j|-,Y,D^{(t-1)},M^{(t-1)},Z^{(t-1)}) \sim CRT(Y_{n_j},\phi^{(t-1)})$

这些变量表示根据中餐厅桌（CRT）分布的（潜在）计数，其定义如下。

如果: $l_j = \sum_{m=1}^{y_j} b_m$，且 $b_m \sim Bernoulli(\phi/(m-1+\phi))$，我们记 $(l_j|-) \sim CRT(y_j,\phi)$ 。

伪代码

Algorithm 1: Gibbs 抽样流程

Input: 观测数据 $\{Y_i\}\ i = 1, 2, \cdots, 400$

Output: 参数后验分布样本 $\theta^{(t)} = (\tilde{\beta}^{(t)}, \phi^{(t)}, \pi_0^{(t)}, p^{(t)}, D^{(t)}, M^{(t)}, Z^{(t)})$

```
Initialize:
目标参数初始化：β̃^(0), φ^(0), π_0^(0), p^(0)
隐含数据初始化：
D_i = 1 if Y_i ≥ 1 else Bernoulli(0.5)        ▷ Y_i 大于 1 的直接为 1，其余部分随机分配 0 或 1
M_i = 1 ~ Bernoulli(0.5)                       ▷ 随机由 Bernoulli 分布分配
Z_i = Y_i if Y_i ≥ 1 else Bernoulli(0.5)      ▷ Y_i 大于 1 的直接赋值，其余部分随机分配 0 或 1
先验参数初始化：μ_β̃, σ²_β̃, e_0, f_0
t ← 1,Iteration,Burn-in                        ▷ 设置迭代次数与 Burn-in 长度
for t = 1, 2, ... ,Iteration do
    (a) 更新 γ_i^(t-1)                          ▷ 需要参数：β̃^(t-1)
        γ_i^(t) = exp(X_i^T β̃^(t)) / (1+exp(X_i^T β̃^(t)))
    (b) 更新 π_0^(t)                            ▷ 利用 R 软件从 Beta 分布抽样得到
        π_0^(t) ~ Beta(n + 1 − Σ_{i=1}^n D_i^(t), 1 + Σ_{i=1}^n D_i^(t))
    (c) 更新 p^(t)                              ▷ 利用 R 软件从 Beta 分布抽样得到
        p^(t) ~ Beta(Σ_{i=1}^n M_i(1 − D_i^(t)) + 1, Σ_{i=1}^n (1 − M_i)(1 − D_i^(t)) + 1)
    (d) 更新隐含样本 D^(t), M^(t), Z^(t)
                     ▷ 利用 P(D_i = a, M_i = b, Z_i = k|Y_i = c, θ^(t-1)) 条件概率抽样得到
        D_i^(t), M_i^(t), Z_i^(t) ~ P(D_i = a, M_i = b, Z_i = k|Y_i = c, θ^(t-1)), i = 1, 2, ..., n
```

伪代码

(e) **更新** $\tilde{\beta}^{(t)}$

方法 1： M-H 抽样法 -> $\tilde{\beta}^{(t)}$

方法 2： Pólya–Gamma 隐变量抽样法 ▷ 利用公式 (14)，通过 BayesLogit 程序包抽样得到

$\omega_i \sim PG((Z_i+\phi^{(t-1)})D_i^{(t-1)}, 0)$，$i = 1, 2, \ldots, n$，记 $\Omega = diag(\omega_1, \omega_2, \cdots, \omega_n)$ ▷ 生成隐变量

计算 $\tilde{H}_\omega = \left(X^T\Omega X + \sigma_{\tilde{\beta}}^{-2} I_s^{-1}\right)^{-1}$,$\tilde{M}_\omega = \tilde{H}_\omega \left(X^T\tilde{\kappa} + \sigma_{\tilde{\beta}}^{-2} I_s^{-1}\mu_{\tilde{\beta}}\right)$,

$\tilde{\kappa} = (\frac{(Z_1-\phi)D_1^{(t-1)}}{2}, \frac{(Z_2-\phi)D_2^{(t-1)}}{2}, \cdots, \frac{(Y_n-\phi)D_n}{2}^{(t-1)})$

$\tilde{\beta}^{(t)}|-, Y, D^{(t-1)}, M^{(t-1)}, Z^{(t-1)}, \omega \sim N\left(\tilde{M}_\omega, \tilde{H}_\omega\right)$ ▷ 正态抽样

(f) **更新** $\phi^{(t)}$

方法 1： M-H 抽样法 -> $\phi^{(t)}$

方法 2： 通过 CRT 的两阶段 Gibbs 抽样 ▷ 利用公式 (15) 抽样得到

step1： $(l_j|-, Y, D^{(t-1)}, M^{(t-1)}, Z^{(t-1)}) \sim CRT(Y_j, \phi^{(t-1)})$

step2： $\gamma_i^{(t-1)} = \frac{\mu_i^{(t-1)}}{\mu_i^{(t-1)} + \phi^{(t-1)}}$

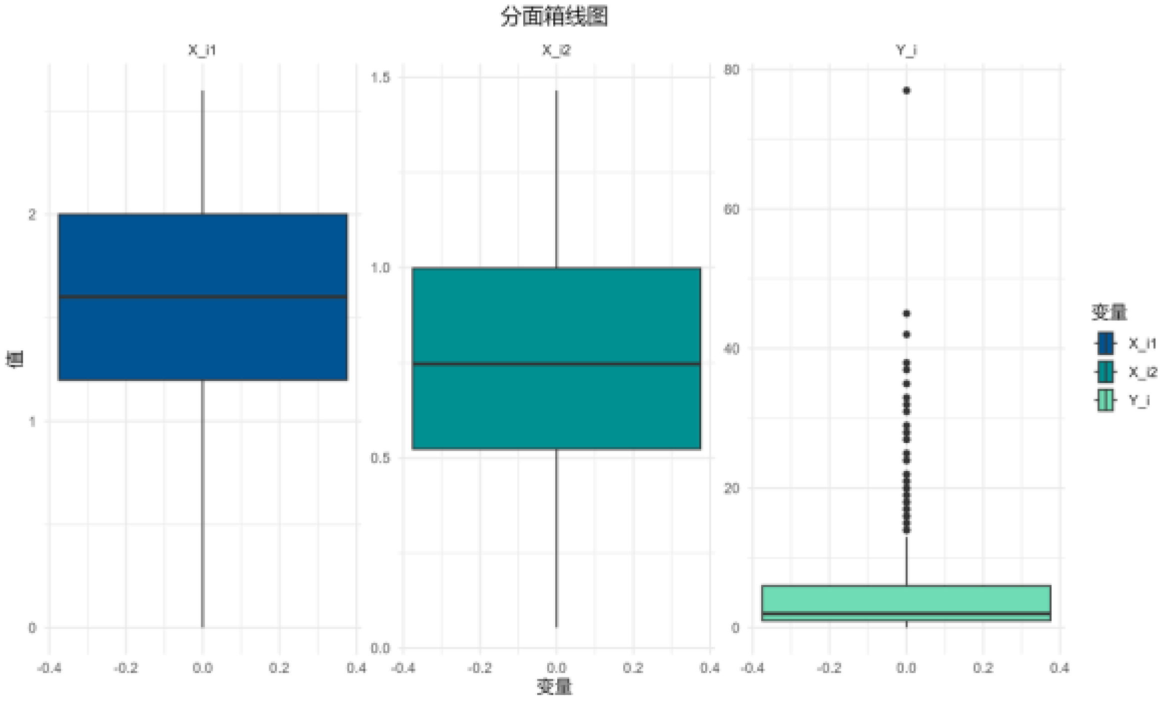
step3： $(\phi^{(t+1)}|-, Y, D^{(t-1)}, M^{(t-1)}, Z^{(t-1)}) \sim Gamma(e_0 + \sum_j l_j, f_0 - \sum_{i=1}^n \log(1-\gamma_i^{(t-1)}))$

end

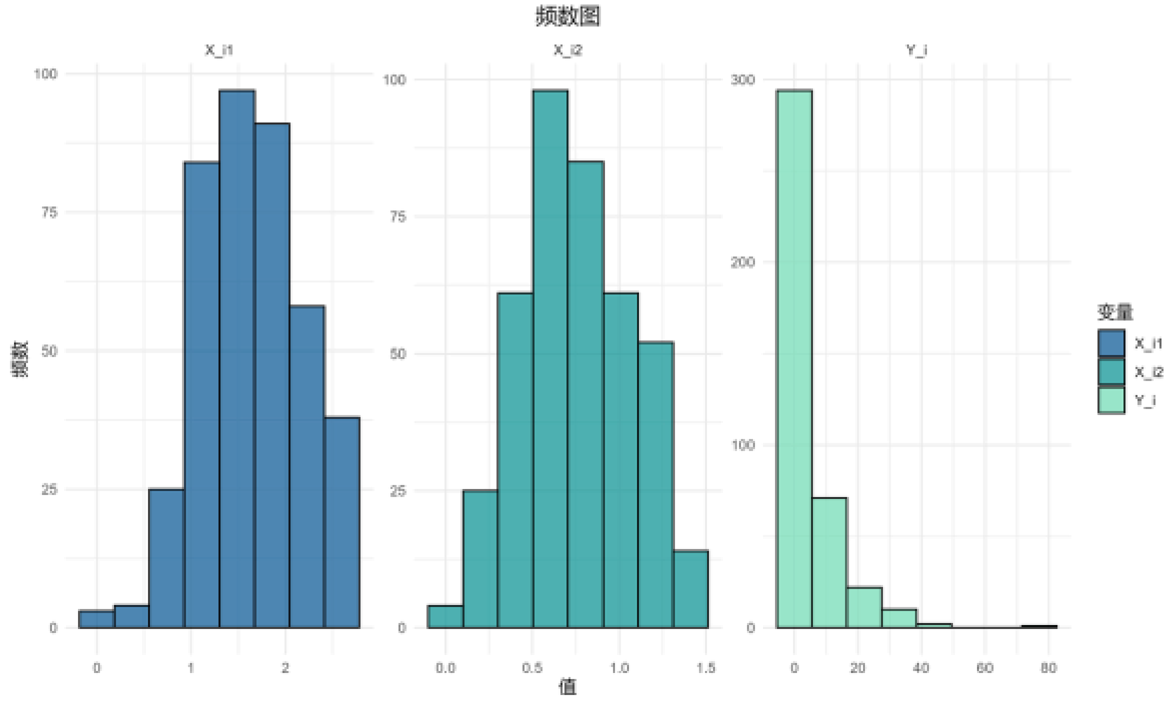
保存去掉 Burn-in 后的数据。

实验结果分析——数据分析与可视化

箱线图



频数图



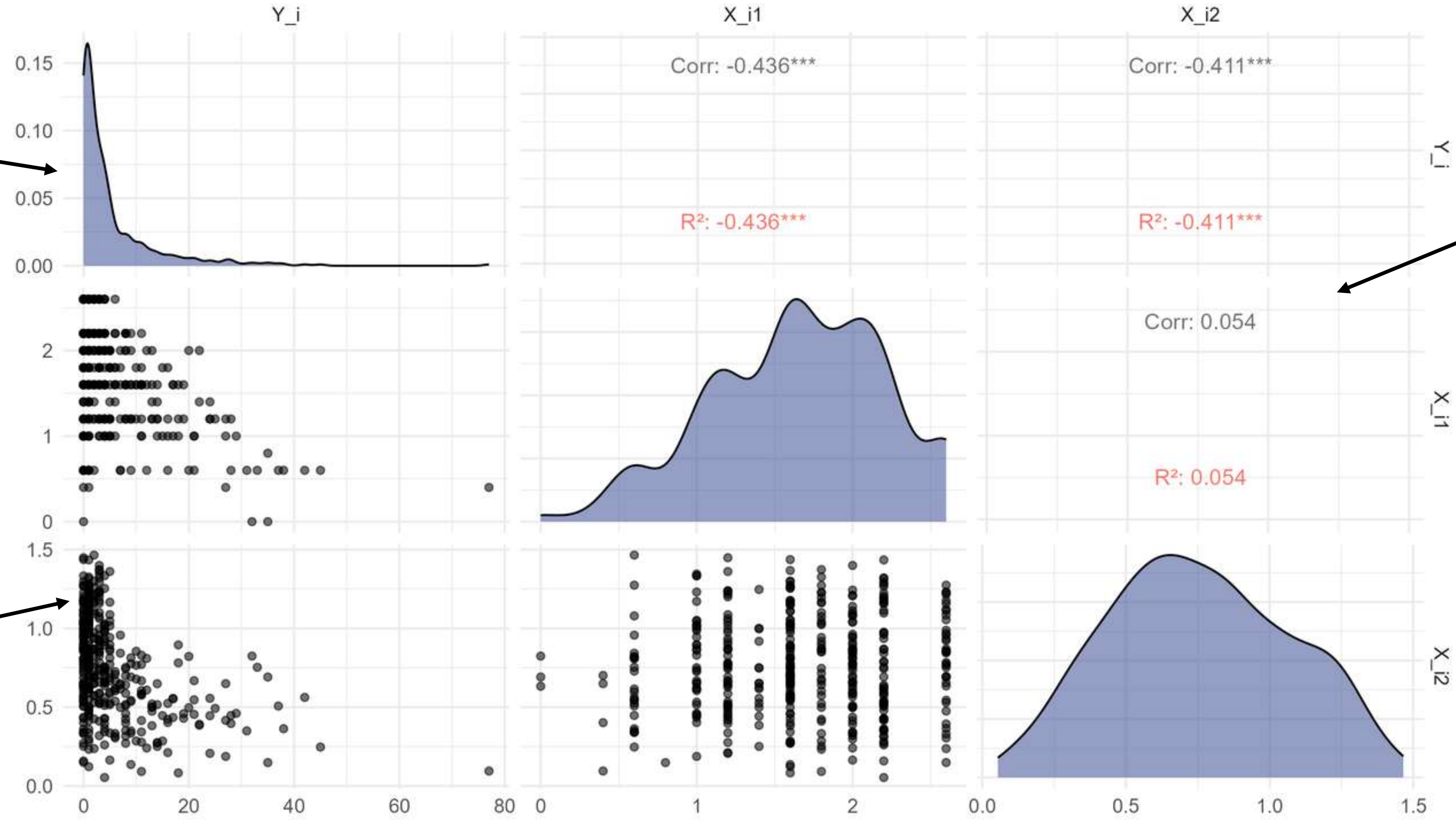
数据的离散分布情况：

- X_{i1}较小一侧在0到1之间的数据比较分散，出现轻微的左偏态；
- X_{i2}的数据均匀地分散在0至1.5之间，近似于正态分布的分布形态；
- Y_i的中位数靠近0，数据存在很多高于正常数据的异常值，出现明显的右偏态。

实验结果分析——数据分析与可视化

相关性图

密度图



相关系数

散点图

生成数据

通过上述分析已经知道，Y_i 与 X_{i1} 及 X_{i2} 均呈现出**负相关关系**，所以在生成数据时将 β_1, β_2 设置为负数；

同时为了保证 $\log(\mu_i) = \log(\frac{\phi\gamma_i}{1-\gamma_i}) = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 = X_i^\top \vec{\beta}$ 合法有效，我们设置较高的 β_0。

为进一步提高估计精度，直接采用 `Data_0-1膨胀负二项回归.xlsx` 数据的 X_{i1} 及 X_{i2} 用于构造测试数据的 Y。基于此，全真实参数设置如下：

β_0	β_1	β_2	ϕ	p	π_0
5.5	-1	-2	4	0.6	0.2

生成数据

生成测试数据Y步骤：

Step1：设定数据和真值

- 用真实样本 X_{i1} , X_{i2} 设定 $X = (1, X_{i1}, X_{i2})$
- 设置全真实参数如上表所示

step2：计算 μ 和 γ

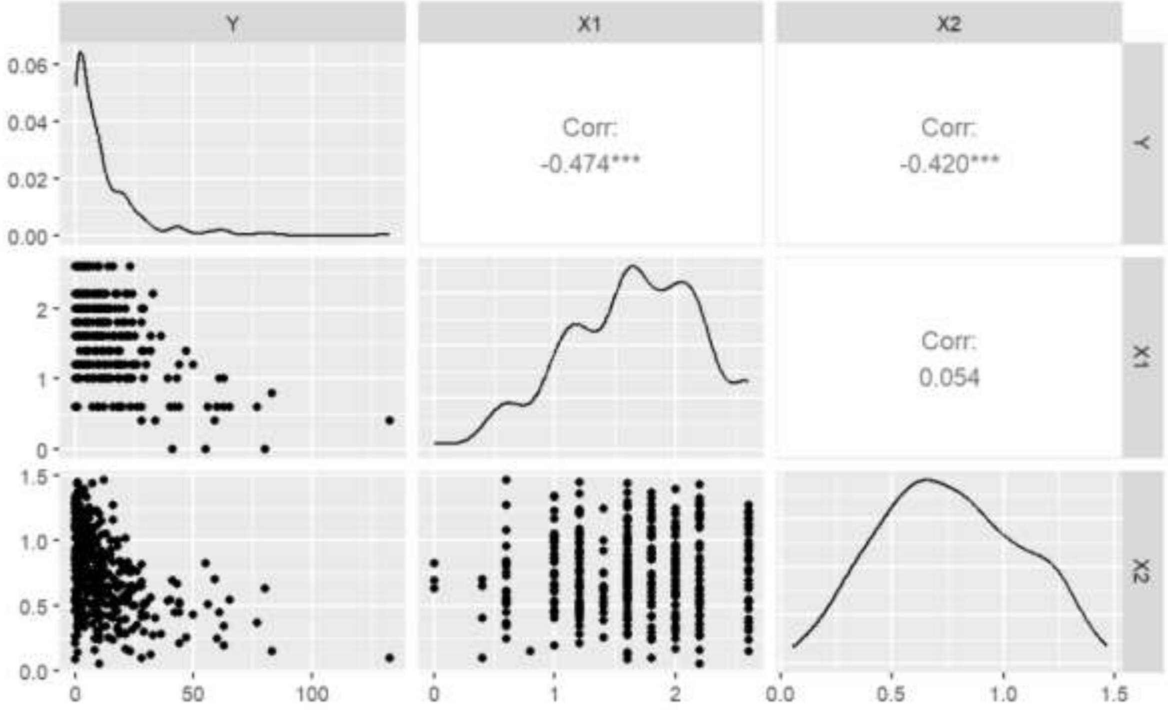
- $\mu = \exp(\mathbf{X}\beta)$
- $\gamma = \mu/(\mu + \phi)$

step3：根据设定参数生成Z，M，D，Y，得测试数据

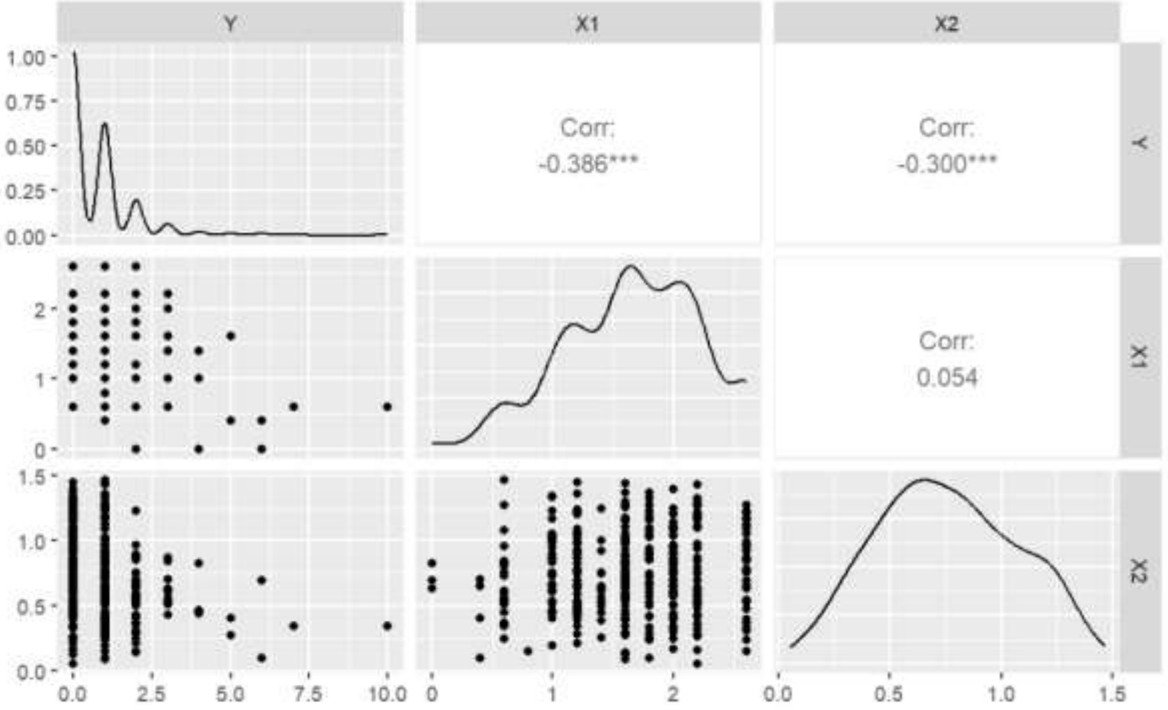
- $Z \sim \mathrm{NB}(n, \phi, \mu)$
- $M \sim \mathrm{B}(n, 1, p)$
- $D \sim \mathrm{B}(n, 1, 1 - \pi_0)$
- $Y = (1 - D) \times M + D \times Z$

生成数据

生成数据的大致分布情况



合适的测试数据



不合适的测试数据

生成数据集与原数据集的**相关情况与分布情况**类似

可作为一个较好的**测试数据集**来评估方法有效性

单参数估计效果测试

在这一部分，我们控制其余参数均为自己设定的真值，只对**目标参数**采用上述的Gibbs抽样方法，来检测估计效果的好坏。

I. 更新隐含数数据D, M, Z

➢ 初始化

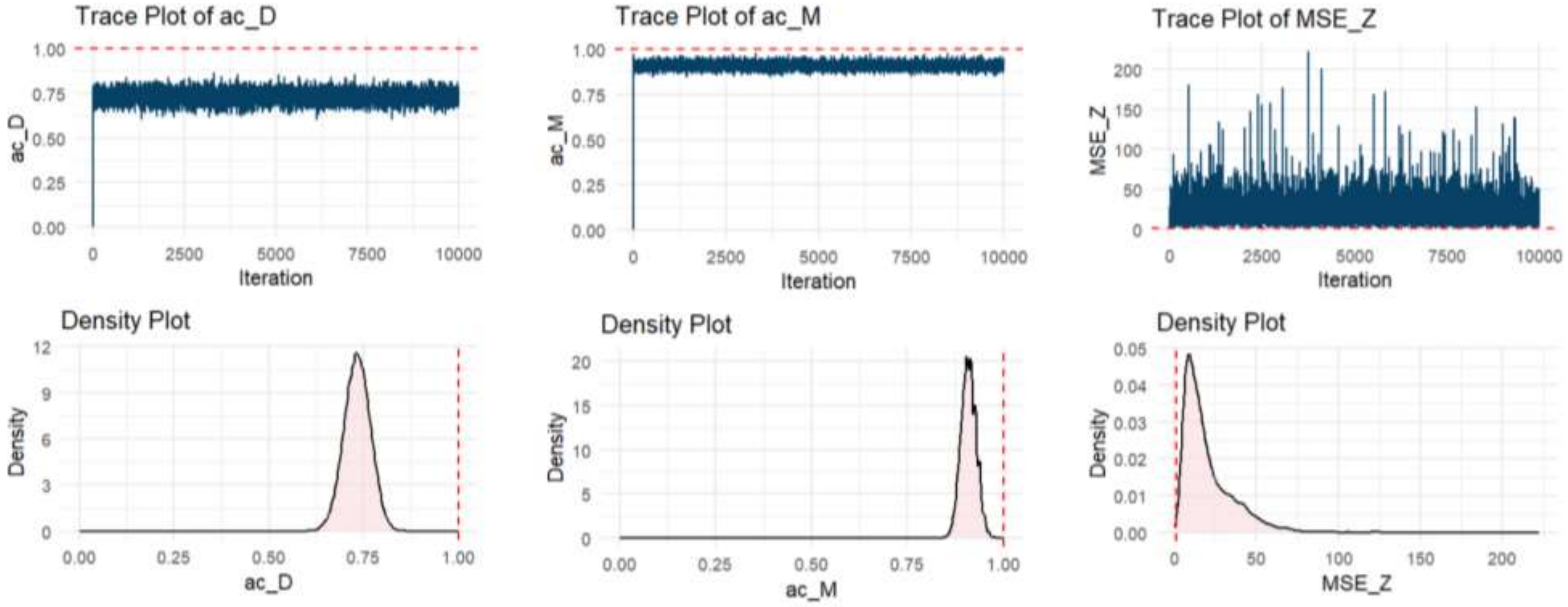
$D_i^{(0)}$	$M_i^{(0)}$	$Z_i^{(0)}$
$=\begin{cases}1 & Y_i >= 1 \\ \sim b(0.5) & else\end{cases}$	$\sim b(0.5)$	$=\begin{cases}Y_i & Y_i >= 1 \\ \sim b(0.5) & else\end{cases}$

- 在10000次Gibbs抽样过程中，重点关注Y=0,1时隐含数据的估计效果。
- 关心的问题：
  - 当 $D_i^{(t)}$ 把样本 Y_i 划分为**负二项分布**时：
    - 估计值 $Z_i^{(t)}$ 和真实值 Z_i 差多少? --> 用**MSE**来衡量差异。
  - 当 $D_i^{(t)}$ 把样本 Y_i 划分为**二项分布**时：
    - 估计值 $M_i^{(t)}$ 和真实值 M_i 差多少? --> 因 M 和 D 取值在0-1，用**准确率**来衡量。
- 监控每次迭代生成的隐含数据 $D^{(t)}, M^{(t)}, Z^{(t)}$ 和真实值的差距，评估参数更新的可靠性。

单参数估计效果测试

➢ 运行结果

隐含数据	更新准确率
$D^{(t)}$	**73.27%**
$M^{(t)}$	**90.94%**
$Z^{(t)}$	**20.24106 (MSE)**



通过分布图可以观察到，对 $D^{(t)}, M^{(t)}$ 更新的**分布较狭长、准确率较稳定**，但对 $Z^{(t)}$ 的估计**有相对较大的偏差。**

即，在该数据因为未知原因仅被简单区分为0、1时，背后实际的计数并不那么容易被估计。

单参数估计效果测试

2. 更新$\tilde{\beta}$

➤ 初始化

iterations	burn_in	$\tilde{\beta}^{(0)}$	$E(\beta_{prior})$	$SE(\beta_{prior})$
10000	4000	(0.2,0.2,0.2)	(4, -1, -2)	(1, 1, 1)

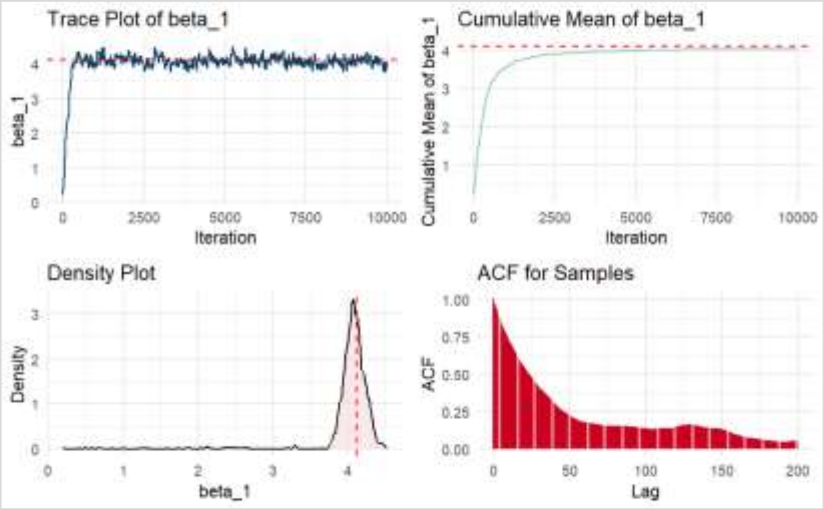
➤ 方法一：M-H抽样

Posterior mean of beta	4.09535	-1.065358	-1.869007
true beta	4.113706	-1	-2

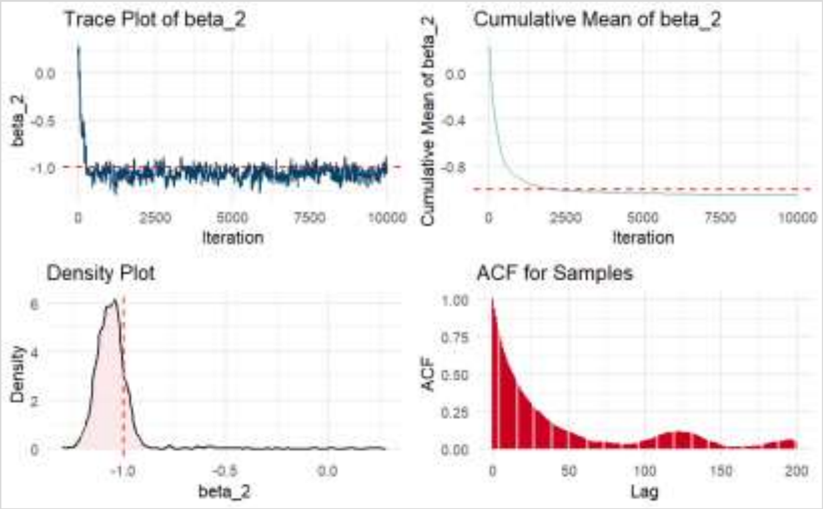
单参数估计效果测试

➢ 方法一：M-H抽样

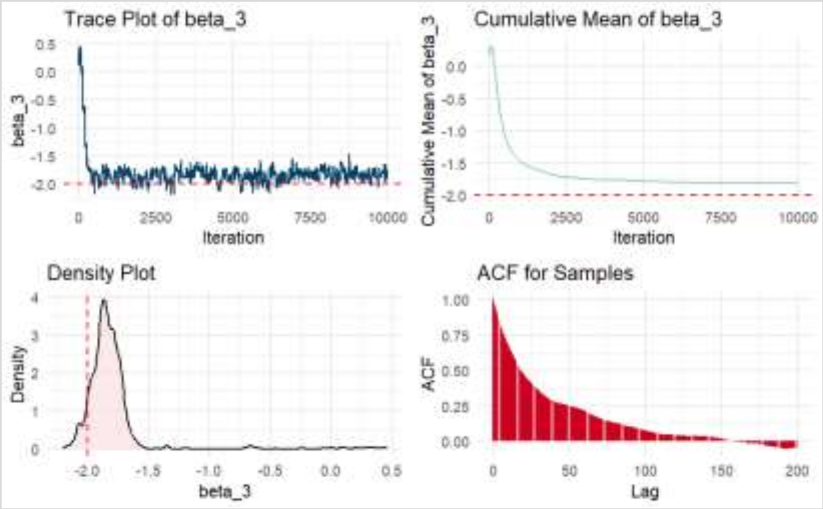
$\tilde{\beta}_0$ 的抽样情况



$\tilde{\beta}_1$ 的抽样情况



$\tilde{\beta}_2$ 的抽样情况



1. **参数收敛性：**
   - 迭代超10000次后均值趋于平稳，收敛值接近真值；
   - burn-in期设置为5000次较为合理。

2. M-H算法存在高拒绝率，导致较高的自相关性。可采用**Thinning策略**改善。

3. 三个参数基本呈现**狭长的正态分布**，说明估计效果良好。

单参数估计效果测试

2. 更新$\tilde{\beta}$

➢ 初始化

iterations	burn_in	$\tilde{\beta}^{(0)}$	$E(\beta_{prior})$	$SE(\beta_{prior})$
10000	4000	(0.2,0.2,0.2)	(4, -1, -2)	(1, 1, 1)

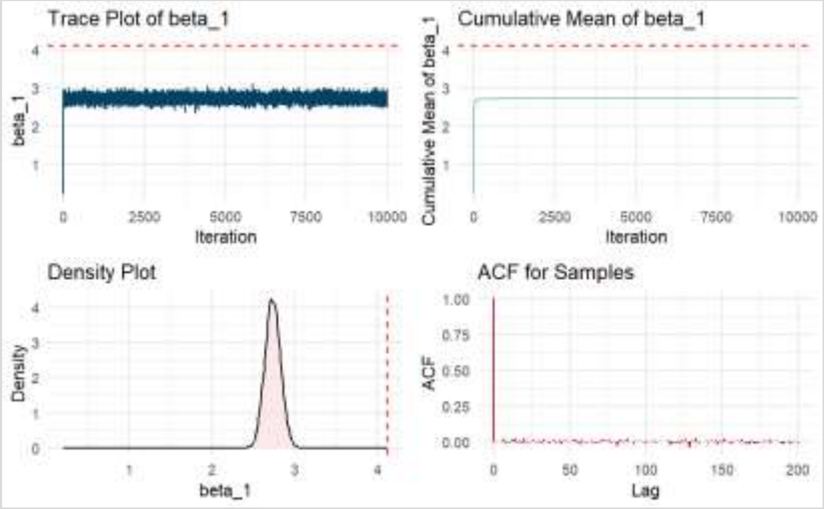
➢ 方法二：Pólya−Gamma隐变量抽样法

Posterior mean of beta	2.728543	-0.6184657	-1.223236
true beta	4.113706	-1	-2

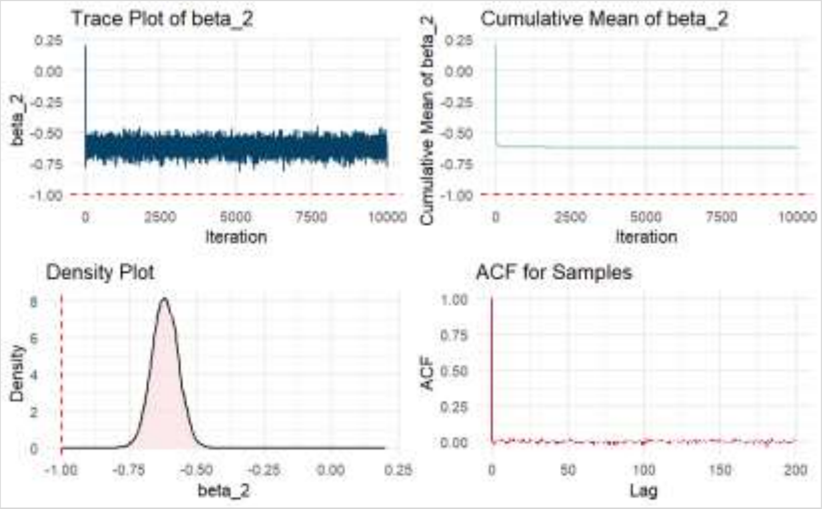
单参数估计效果测试

➢ 方法一： Pólya−Gamma隐变量抽样法

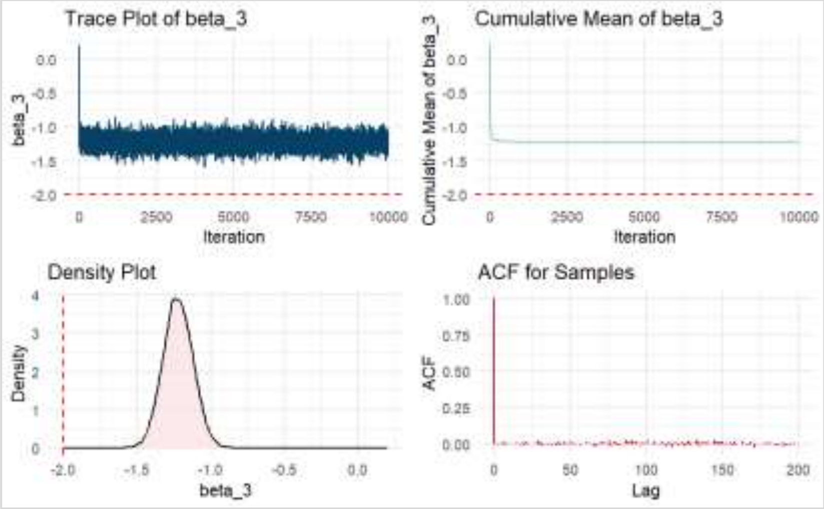
$\tilde{\beta}_0$ 的抽样情况



$\tilde{\beta}_1$ 的抽样情况



$\tilde{\beta}_2$ 的抽样情况



1. **方法二：**
   - 无高拒绝率问题；
   - 大样本下**矩阵求逆**计算开销大——以计算复杂度换取时间效率。

2. 该方法表现出**稳定的极度有偏现象**，推测复现过程可能缺失关键实现细节。

➢ **综上，我们决定后续Gibbs抽样继续采用M-H采样。**

单参数估计效果测试

3. 更新ϕ

➢ 初始化

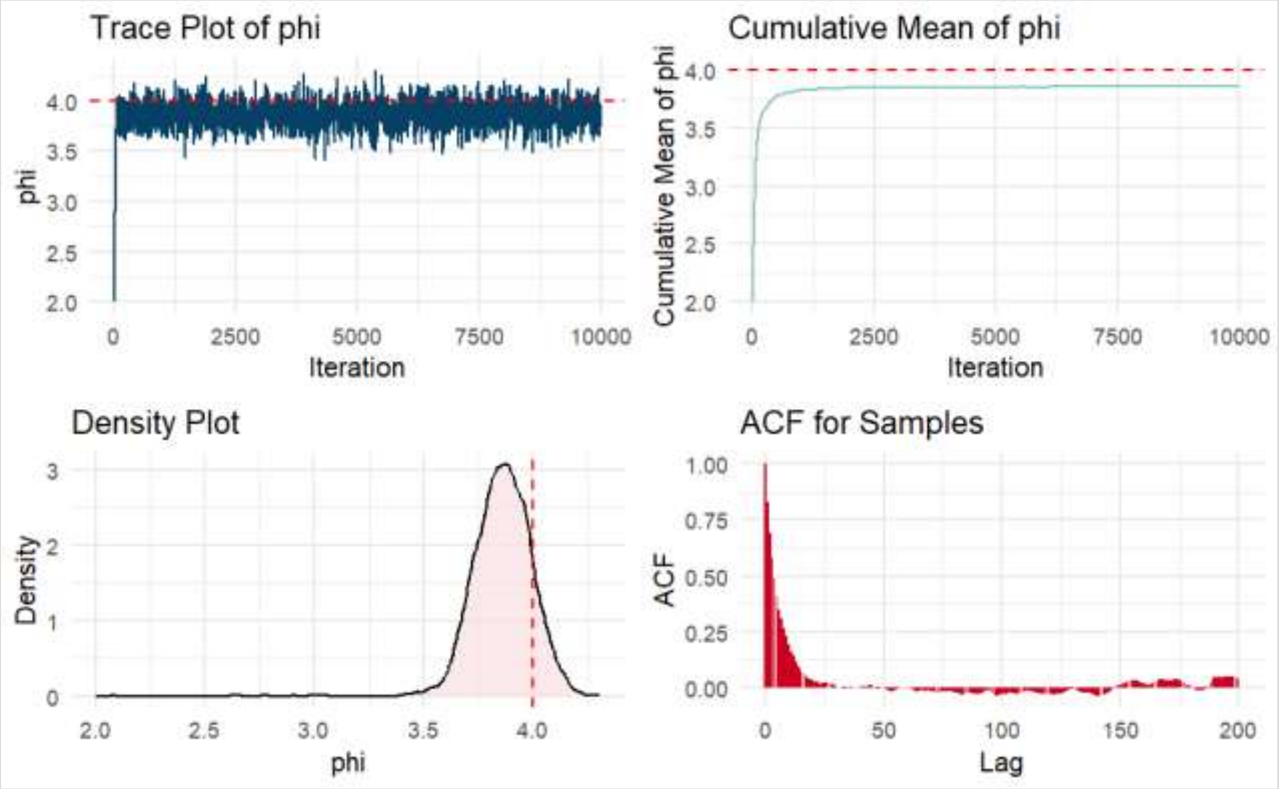
iterations	burn_in	$\phi^{(0)}$	e_0	f_0
10000	4000	2	12	2

➢ 方法一：M-H抽样

➢ 方法二：通过CRT的两阶段Gibbs抽样生成ϕ

单参数估计效果测试

➢ 方法一： M-H抽样



ϕ的抽样情况

➢ **估计后验分布均值：3.864**

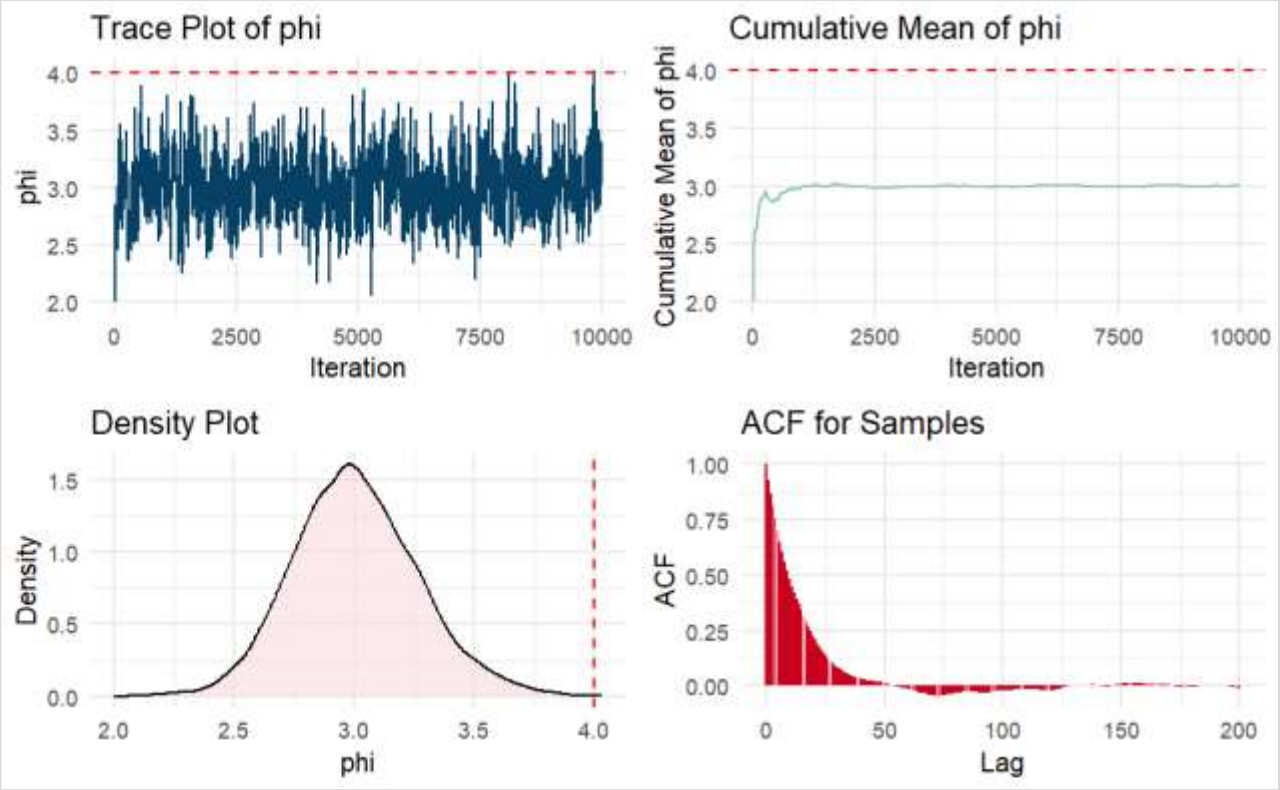
1. 5000次迭代后均值基本收敛
   --> 建议burn-in：2000~4000
2. 多次尝试后发现，估计是否收敛于真实值受数据特性影响。

3. 依旧存在问题：
   - 高拒绝率导致难收敛；
   - 存在明显自相关性。

在后续完整Gibbs抽样中上述问题更显著。

单参数估计效果测试

➢ 方法二：通过CRT的两阶段Gibbs抽样生成ϕ



ϕ的抽样情况

➢ 估计后验分布均值：3.005026

该方法同样表现出**极度有偏。**

推测其原因在于该方法在处理“**共用相同过离散参数ϕ，拥有不同概率参数γ**”的模型时，可能并不具备良好的适用性。

➢ 综上，我们决定后续Gibbs抽样继续采用M-H算法实现对ϕ的采样。

全参数Gibbs抽样

➢ 初始化

$\tilde{\beta}^{(0)}$	$\phi^{(0)}$	$\pi_0^{(0)}$	$p^{(0)}$	$D_i^{(0)}$	$M_i^{(0)}$	$Z^{(0)}$
(0.2, -0.2, -0.2)	2	0.5	0.5	$=\begin{cases}1 & Y_i >= 1 \\ \sim b(0.5) & else\end{cases}$	$\sim b(0.5)$	$=\begin{cases}Y_i & Y_i >= 1 \\ \sim b(0.5) & else\end{cases}$

➢ 超参数设置

1. 先验分布设置:

$E(\beta_{prior})$	$SE(\beta_{prior})$	e_0	f_0	σ_0^2
(1,-1,-1)	(1,1,1)	12	2	0.01

- $E(\beta_{prior})$: β 先验分布的均值
- $SE(\beta_{prior})$: β 先验分布的标准差
- e_0 ： ϕ 先验分布（Gamma 分布）形状参数
- f_0 : ϕ 先验分布（Gamma 分布）速率参数

2.提议分布设置:

➢ M-H算法提议函数 $q(x, x')$ 设为 $N(x, \sigma_0^2)$

全参数Gibbs抽样

➢ 实验设置与估计结果:

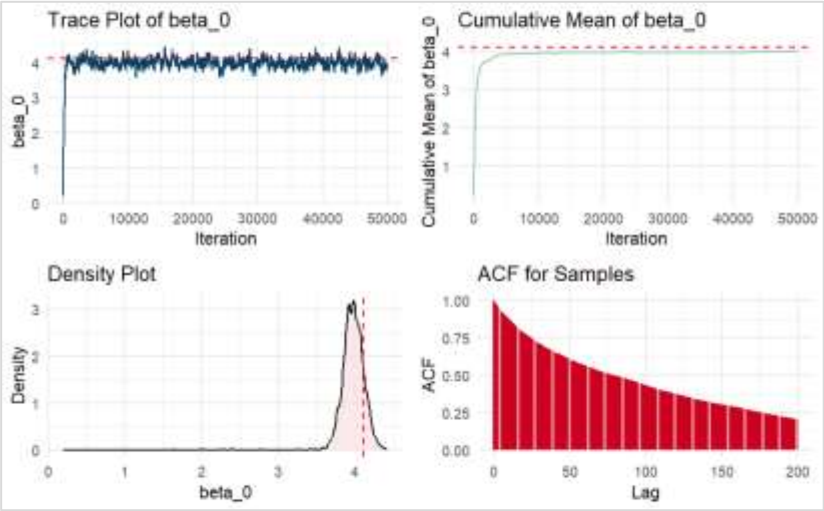
- 迭代次数：50000　　　　burn-in ：10000

参数	β_0	β_1	β_2	ϕ	p	π_0
真实值	5.5	-1	-2	4	0.6	0.2
Gibbs抽样估计后验分布均值:	5.499993	-1.005951	-2.054578	4.572603	0.6253828	0.2075729

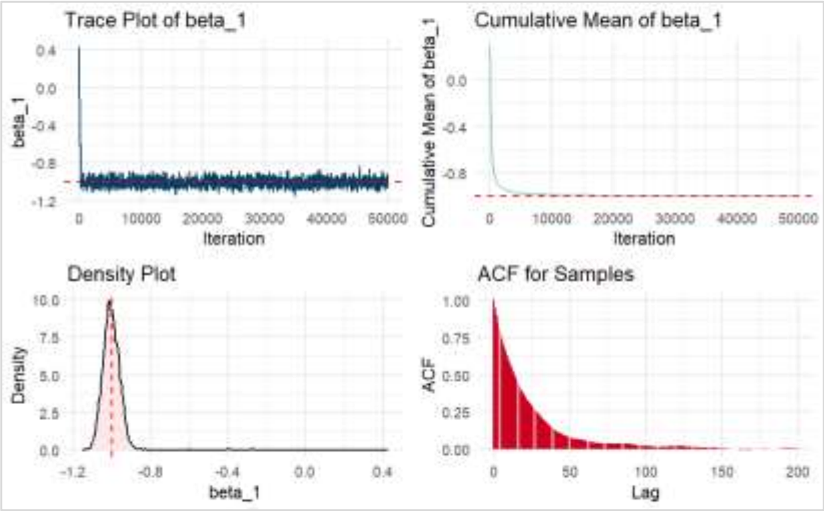
全参数Gibbs抽样

➢ 实验结果： 1. 回归系数$\vec{\boldsymbol{\beta}}$

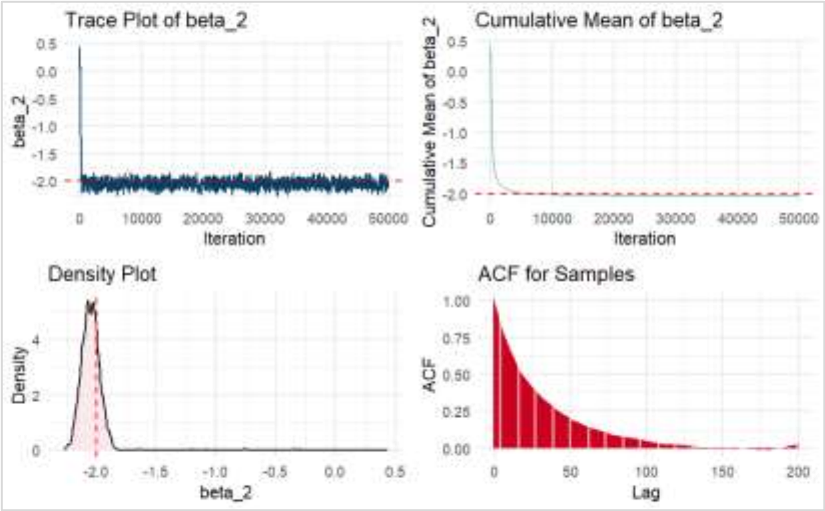
$\tilde{\beta}_0 = \beta_0 - \log(\phi)$



β_1



β_2

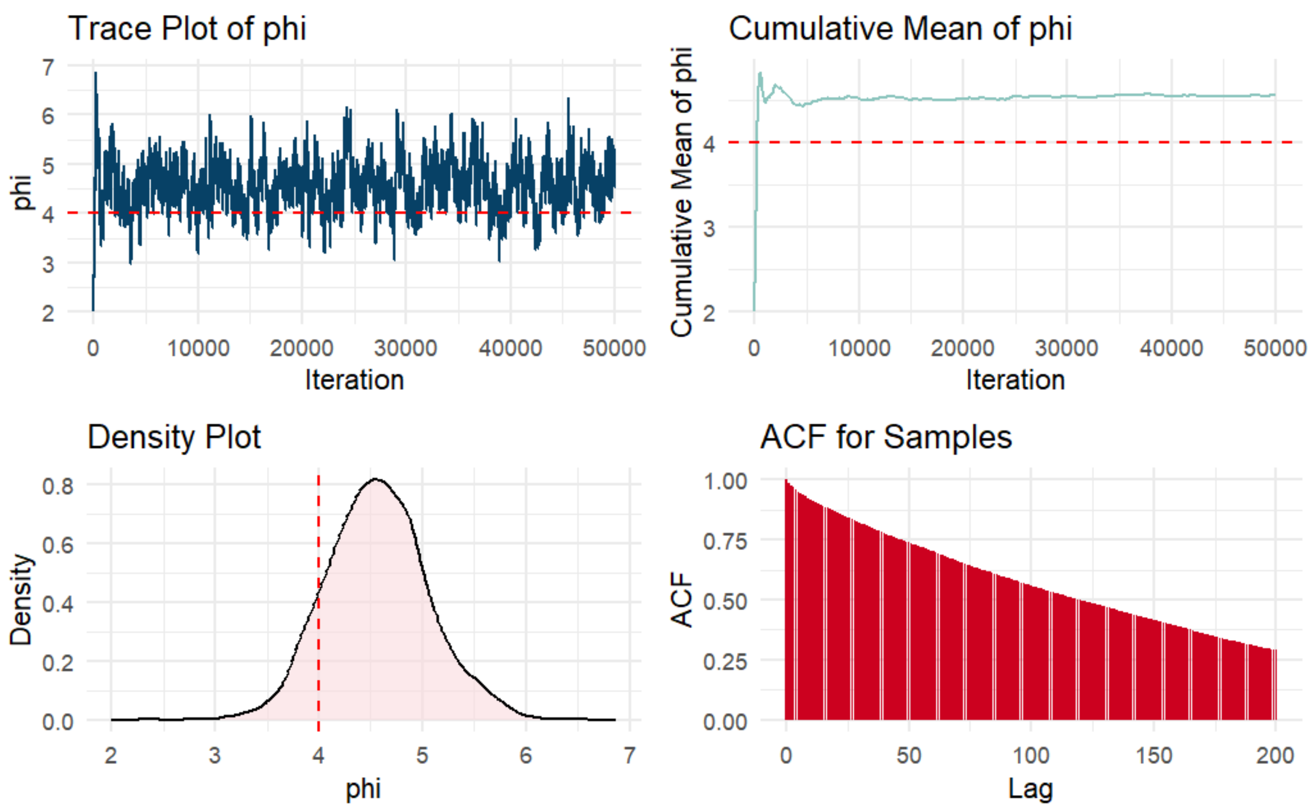


➢ 三个回归系数的估计基本已收敛到真实参数。

然而存在一个奇怪的现象：通过 $\tilde{\beta}_0 = \beta_0 - \log(\phi)$ 转换后得到的 β_0 估计是几乎无偏的。但由于后续会发现 ϕ 的估计存在偏差，导致 $\tilde{\beta}_0 = \beta_0 - \log(\phi)$ 仍然存在偏倚。

全参数Gibbs抽样

➢ 实验结果： **2. 过离散参数** ϕ：$Z_i \sim NB(\mu_i, \phi)$

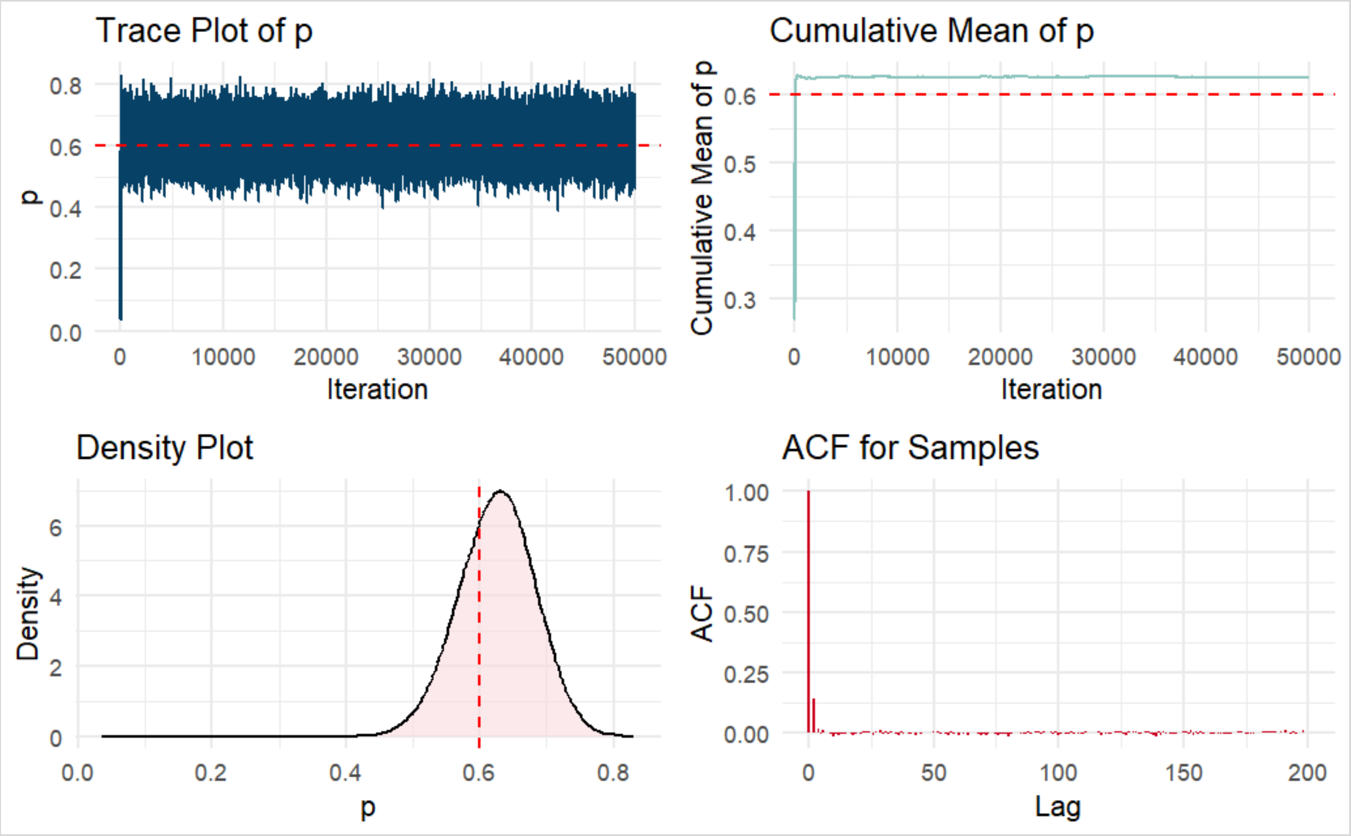


➢ ϕ **的估计出现了较大的偏移。**

- 从轨迹图来看，ϕ 的**收敛情况较差**，但累积均值图已经基本收敛。猜测更多的迭代次数也很难再有更多精度的改善。
- 在当前迭代过程中，ϕ 的**自相关性**在所有参数中最为严重。

全参数Gibbs抽样

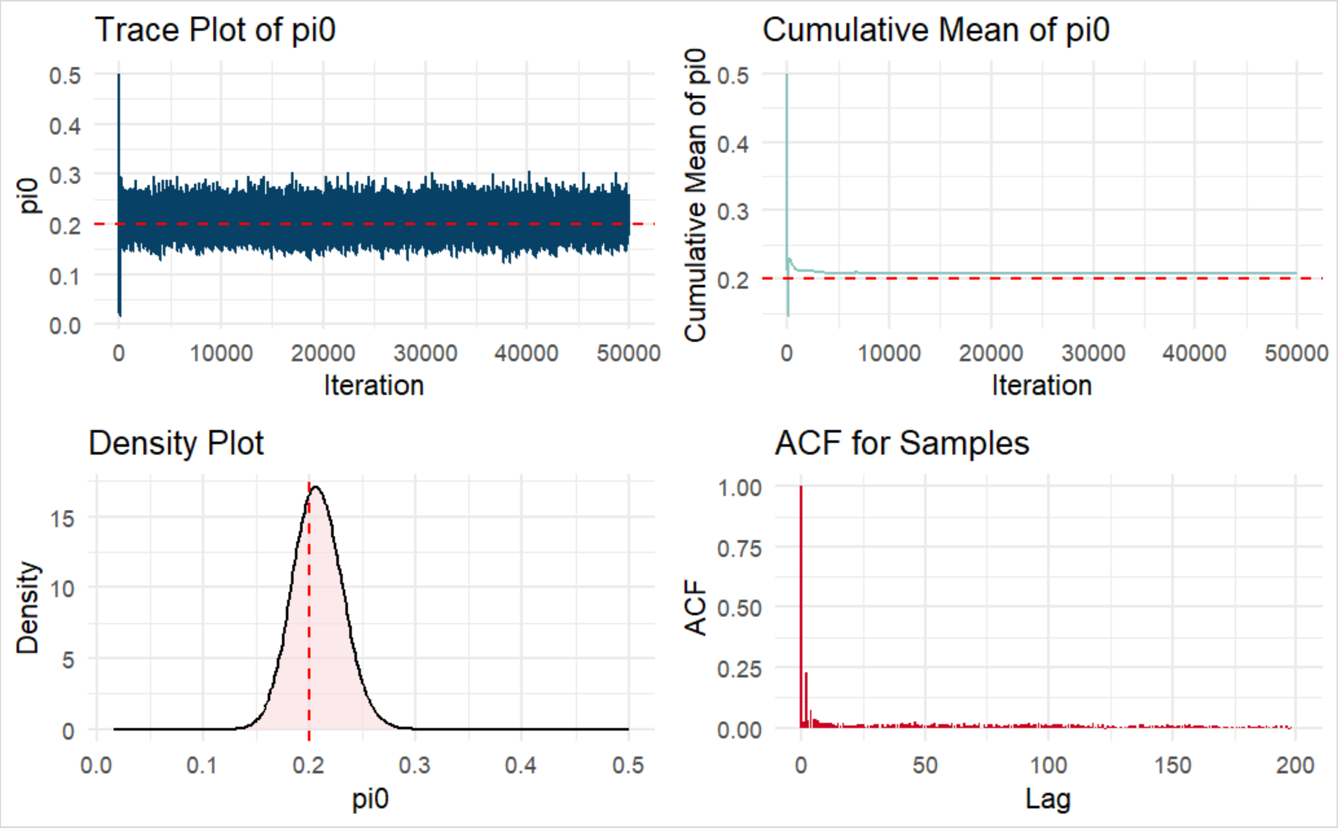
➤ 实验结果： **3.二项分布参数** p： $M_i \sim B(1,p)$



- 轨迹图显示**良好的收敛态势**，**自相关性**不高，形成的后验分布**非常接近完美的正态形式**。
- 5000次迭代后，累积均值已经基本收敛，在该样本下，进一步的估计已难以带来显著改善。

全参数Gibbs抽样

➢ 实验结果： **4. 数据混合来源于二项分布的概率** π_0： $D_i \sim B(1,1-\pi_0)$



- 基本结论与对参数p的估计一致。
- 这也表明，**具有明确且易于抽样的满条件分布形式**对估计结果的准确性和效率具显著优势。

未知参数数据估计

基础设置和上面过程基本一样，只是额外修改了先验分布的超参数。

- 初始化

$\tilde{\beta}^{(0)}$	$\phi^{(0)}$	$\pi_0^{(0)}$	$p^{(0)}$	$D_i^{(0)}$	$M_i^{(0)}$	$Z^{(0)}$
(0.2, -0.2, -0.2)	2	0.5	0.5	$= \begin{cases} 1 & Y_i >= 1 \\ \sim b(0.5) & else \end{cases}$	$\sim b(0.5)$	$= \begin{cases} Y_i & Y_i >= 1 \\ \sim b(0.5) & else \end{cases}$

- 超参数设置

$E(\beta_{prior})$	$SE(\beta_{prior})$	e_0	f_0	σ_0^2
(4,-1,-2)	(0.16,0.05,0.08)	12	2	0.01

为什么修改为这样的超参数?
- 通过预先进行的估计结果大致猜测，有助于模型更快收敛。

未知参数数据估计

➢ 实验设置与估计结果:

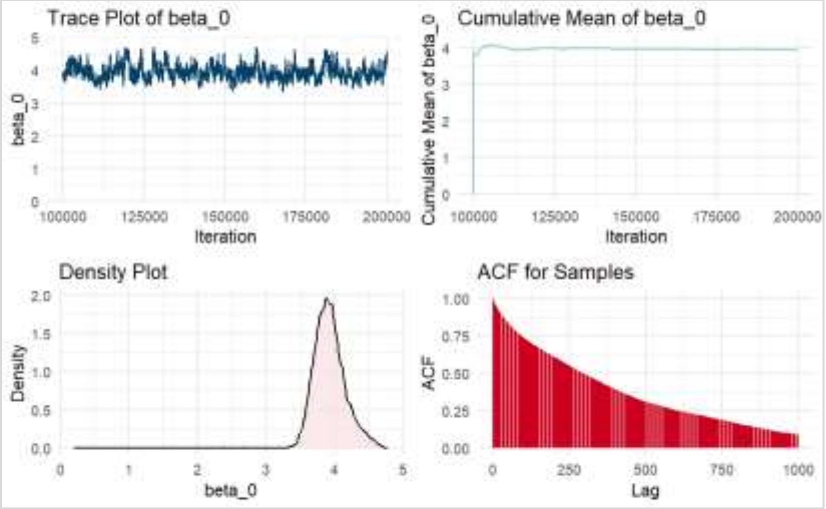
- 迭代次数：20万次　　　　burn-in ：10万次

参数	β_0	β_1	β_2	ϕ	p	π_0
Gibbs抽样估计后验分布均值:	4.867892	-1.083863	-2.120212	2.521978	0.4746306	0.2100464
猜测真实参数	5	-1	-2	2.5	0.48	0.2

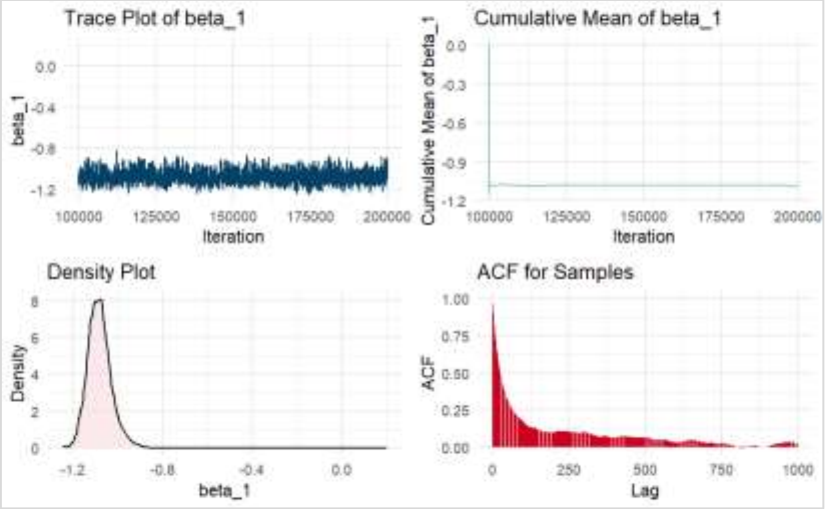
顺手做一个猜测

未知参数数据估计结果

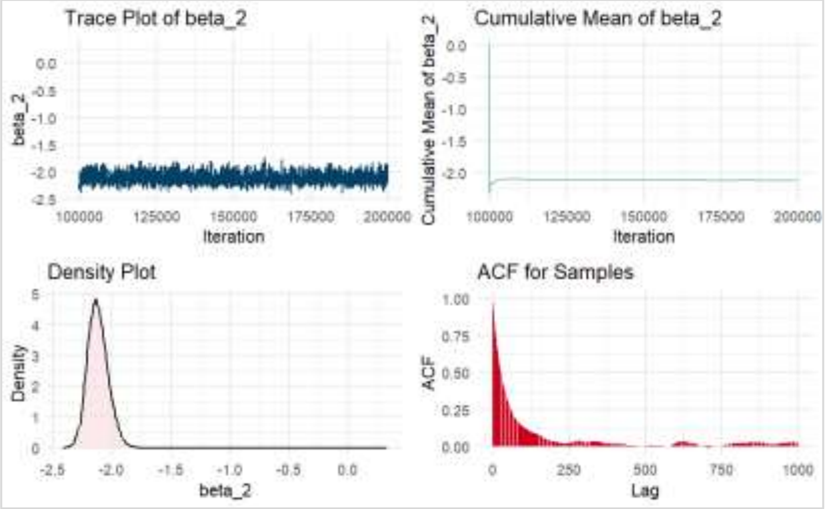
$\tilde{\beta}_0 = \beta_0 - \log(\phi)$



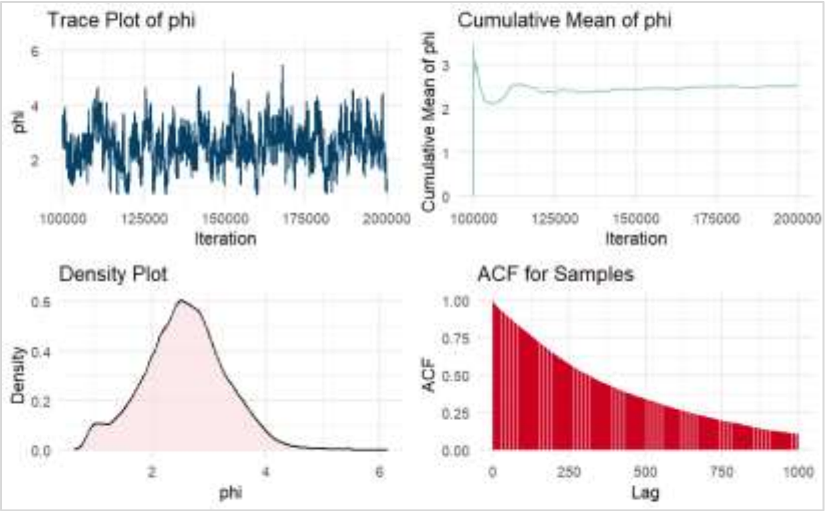
β_1



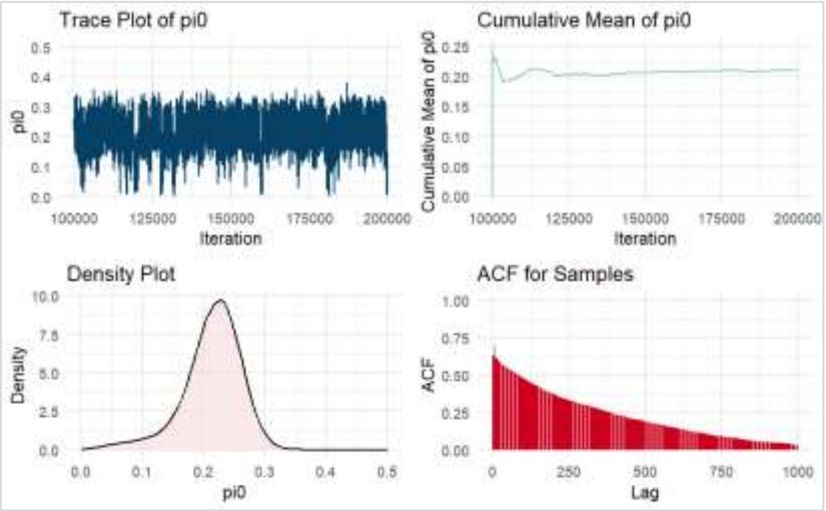
β_2



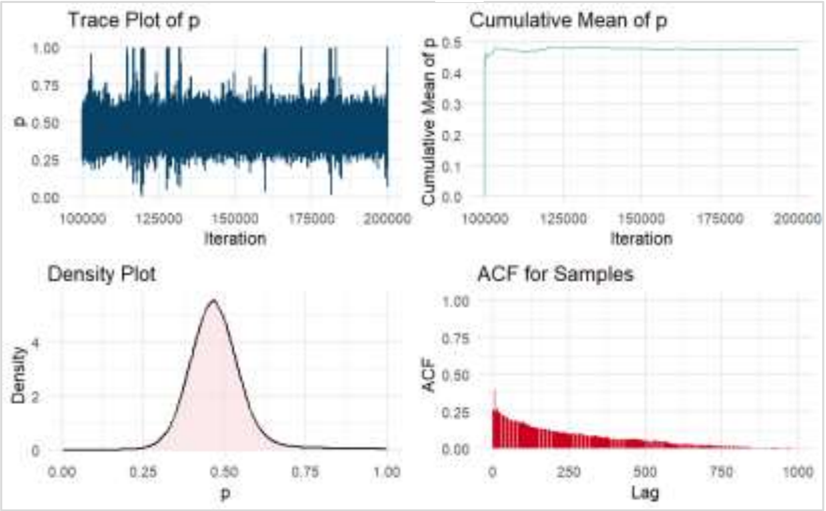
ϕ



π_0



p



感谢观看

GROUP 4

做这个作业真的超努力的第四组