

作业三

强烈推荐阅读方式（更完善的目录导航与排版）：

飞书在线阅读文档：https://owj4v466fmb.feishu.cn/docx/H7AkdrJZ1oSpLnx7RZzcOoyMn9b?from=from_copylink

作业三、0-1膨胀负二项回归的参数估计问题

问题描述及与分析

1. 问题描述

在该问题的研究中，我们探讨**0-1膨胀负二项回归模型的参数估计问题**。数据集 `Data_0-1膨胀负二项回归.xlsx` 包含 $n = 400$ 个独立同分布样本点 $(Y_i, X_{i1}, X_{i2})(i = 1, 2, \dots, n)$ ，其中 Y_i 是响应变量， X_{i1} 和 X_{i2} 是自变量。问题的主要目标是对6个参数 $(\beta_0, \beta_1, \beta_2, \phi, \pi_0, p)$ （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 ）进行估计，并分析其对响应变量的影响。

2. 问题分析

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0/1值的情况，这些0/1值可能是“**真实的零/一**”（实际计数为0/1）和“**伪零/一**”（由于其他因素导致计数未发生或被简单区分为1），为了拟合这种情形的数据，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, \dots, Y_n) 实际上是由潜在序列 (M_1, \dots, M_n) 和 (Z_1, \dots, Z_n) 根据隐变量序列 (D_1, \dots, 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。

模型理论推导

基于上述描述，可以先写出模型的一般形式：

1. 模型与数据

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


2. 似然函数

- 完全数据的单样本概率函数可以写为：

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

- 完全数据的似然函数：

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

 为了写出上式，我们一一来看其中的部分。

1. D_i 的概率容易写出：

$$p(D_i) = (1 - \pi_0)^{D_i} \pi_0^{1-D_i} \quad (2)$$

2. 观测数据的条件概率函数 $p(Y_i, M_i, Z_i | D_i, \theta)$ 稍复杂，需要用特殊的限制来将 $D_i = 0, 1$ 两种情况均写出在一个式子内，如何构造这样的形式？

首先回顾**无条件概率**的形式，并用**全概率公式**将其写出：

$$\begin{aligned} P(Y_i, M_i, Z_i | \theta) &= P(D_i = 1 | \theta) P(Y_i, M_i, Z_i | D_i = 1, \theta) + P(D_i = 0, M_i, Z_i | \theta) P(Y_i | D_i = 0, \theta) \\ &= (1 - \pi_0) P(Z_i | \mu, \phi) + \pi_0 P(M_i) \\ &= (1 - \pi_0) \frac{\Gamma(\phi + Z_i)}{\Gamma(Z_i + 1) \Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi} \right)^{Z_i} \left(\frac{\phi}{\mu_i + \phi} \right)^\phi + \pi_0 p^{M_i} (1 - p)^{1-M_i} \end{aligned}$$

注意，当 $D_i = 0$ 时 Y_i 只能取 0, 1，在上式中 $P(Y_i|D_i = 0, \theta)$ 是自然隐含了 $D_i = 0$ 对 Y_i 的限制，但在后续迭代要用 Y_i 反过来推断 D_i 时就要特意地考虑这个反过来的条件，从而优化估计的效果。

参照上面这个形式，仍用 $1 - D_i$ ， D_i 来控制两部分的取舍，就可以写出 $p(Y_i|D_i, \theta)$ ：

$$p(Y_i, M_i, Z_i|D_i, \theta) = \left[\frac{\Gamma(\phi + Z_i)}{\Gamma(Z_i + 1)\Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi} \right)^{Z_i} \left(\frac{\phi}{\mu_i + \phi} \right)^\phi \right]^{D_i} [p^{M_i}(1-p)^{1-M_i}]^{(1-D_i)} \quad (3)$$

为了便于后续讨论，我们还需要介绍一个trick来将替换一下上式的几个参数形式。

回顾伯努利实验“第 r 次试验成功时试验次数”定义，直接导出定义的负二项分布：

$X \sim Nb(r, p)$, $r \in \mathbb{Z}^+$, $p \in [0, 1]$ 是单次试验失败概率，其概率分布如下：

$$P(X = x) = \binom{x-1}{r-1} p^{x-r} (1-p)^r = \frac{\Gamma(x)}{\Gamma(r)\Gamma(x-r+1)} p^{x-r} (1-p)^r, \quad x = r, r+1, \dots$$

此定义下随机变量 X 并不从 0 开始，为了将其转化为一般形式所以构造平移变换 $Z = X - r$ ，此时

$$P(Z = z) = \frac{\Gamma(z+k)}{\Gamma(r)\Gamma(z-r)} p^z (1-p)^r, \quad y = 0, 1, \dots$$

对比原题中概率分布列形式：

$$P(Z_i = z_i | \mu_i, \phi) = \frac{\Gamma(\phi + z_i)}{\Gamma(z_i + 1)\Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi} \right)^{z_i} \left(\frac{\phi}{\mu_i + \phi} \right)^\phi, \quad z_i = 0, 1, 2, \dots$$

可知参数有对应关系：

- $\mu_i = r \frac{p_i}{1-p_i}$
- $\phi = r$

为了与 $M_i \sim B(1, p)$ 的参数区分，我们记 $p_i = \gamma_i$ ，并将参数转换形式为 $\mu_i = \phi \frac{\gamma_i}{1-\gamma_i}$ ，这样一来 (3) 式中

- $\frac{\mu_i}{\mu_i + \phi} = \gamma_i, \frac{\phi}{\mu_i + \phi} = 1 - \gamma_i$

而 μ_i 相对辅助变量 X_{i1}, X_{i2} 的回归形式转换为：

$$\log(\mu_i) = \log\left(\frac{\phi\gamma_i}{1-\gamma_i}\right) = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 = X_i^\top \vec{\beta} \quad (4)$$

其中 $\vec{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ ， $X_i = (1, X_{i1}, X_{i2})$ 。

对式 (4) 稍做变形即可得到 $\gamma_i = \frac{\exp(X_i^\top \tilde{\beta})}{1 + \exp(X_i^\top \tilde{\beta})}$ ，其中，

$\tilde{\beta}_0 = \beta_0 - \ln \phi$ ， $\tilde{\beta}_j = \beta_j$ ， $(j = 1, 2)$ ，将此式子代入式 (3) 即可得到：

$$p(Y_i, M_i, Z_i | D_i, \theta) = \left[\frac{\Gamma(\phi + Z_i)}{\Gamma(Z_i + 1)\Gamma(\phi)} \right]^{D_i} \frac{\{\exp(X_i^T \tilde{\beta})\}^{Z_i D_i}}{\{1 + \exp(X_i^T \tilde{\beta})\}^{(Z_i + \phi) D_i}} \left[p^{M_i(1-D_i)} (1-p)^{(1-M_i)(1-D_i)} \right] \quad (5)$$

3. 先验分布

为了得到目标参数后验分布的正比形式，我们需要先写出参数的先验分布，然后跟上面的完全数据似然函数拼起来。

对于参数： $(\beta_0, \beta_1, \beta_2, \phi, \pi_0, p)$ ，我们一一确定它们的先验。

假设：

- 参数向量 $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)^\top$ 先验分布建议为多维正态分布[1],[2]，即 $\tilde{\beta} \sim N_3(\mu_{\tilde{\beta}}, \sigma_{\tilde{\beta}}^2 I_3)$ ，其中 $\mu_{\tilde{\beta}}$ 是已知向量， $\sigma_{\tilde{\beta}}^2$ 是已知常数。
- 参数 ϕ 先验分布建议为Gamma分布[3]，即 $\phi \sim \text{Gamma}(e_0, f_0)$ ，其中 e_0 是形状参数， f_0 是速率参数
- 参数 π_0 先验分布建议为区间 $[0, 1]$ 上的均匀分布[2]，即 $\pi_0 \sim U(0, 1)$ 。
- 参数 p 先验分布建议为区间 $[0, 1]$ 上的均匀分布，即 $p \sim U(0, 1)$ 。

进一步假设 $\tilde{\beta}, \phi, \pi_0$ 和 p 相互独立，则 $(\tilde{\beta}, \phi, \pi_0, p)$ 的联合先验分布

$\pi(\tilde{\beta}, \phi, \pi_0, p) = \pi(\tilde{\beta})\pi(\phi)\pi(\pi_0)\pi(p)$ 。结合式(1)(2)(5)，得到数据扩充下 $(\tilde{\beta}, \phi, \pi_0, p)$ 的**后验分布正比形式为**：

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

由式（6）即可得到**各参数的满条件分布**（下式中"－"表示其余给定参数）：

$$\bullet \quad \pi(\tilde{\beta} | -, Y, D, M, Z) \propto \prod_{i=1}^n \frac{\{\exp(X_i^T \tilde{\beta})\}^{Z_i D_i}}{\{1 + \exp(X_i^T \tilde{\beta})\}^{(Z_i + \phi) D_i}} \times \pi(\tilde{\beta}) \quad (7)$$

$$\bullet \quad \pi(\phi | -, Y, D, M, Z) \propto \prod_{i=1}^n \left[\frac{\Gamma(\phi + Z_i)}{\Gamma(\phi)} \right]^{D_i} \times \prod_{i=1}^n \{1 + \exp(X_i^T \tilde{\beta})\}^{-\phi D_i} \times \pi(\phi) \quad (8)$$

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

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

前面已经提到，观测数据 $\{Y_i\}$ 本质是隐含的 $\{Z_i\}$ 和 $\{M_i\}$ 构成的数据，加上混合来源标签 $\{D_i\}$ 均需要在后续Gibbs抽样中迭代更新，考虑具体更新规则时先指出 Y_i 观测到不同数值时背后的事件组合[4]：

- 情况1： $D_i = 0, M_i = 0$ ：

该情况的概率为 $P(D_i = 0, M_i = 0) = \pi_0(1-p)$ ，此时 $Y_i = (1-0)0 + 0Z_i = 0$

- 情况2： $D_i = 0, M_i = 1$ ：

该情况的概率为 $P(D_i = 0, M_i = 1) = \pi_0 p$ ，此时 $Y_i = (1-0)1 + 0Z_i = 1$

- 情况3： $D_i = 1, M_i = 0$ ：

该情况的概率为 $P(D_i = 1, M_i = 0) = (1 - \pi_0)(1-p)$ ，此时 $Y_i = (1-1)0 + 1Z_i = Z_i$

- 情况4： $D_i = 1, M_i = 1$ ：

该情况的概率为 $P(D_i = 1, M_i = 1) = (1 - \pi_0)p$ ，此时 $Y_i = (1-1)1 + 1Z_i = Z_i$

由于 Z_i 是负二项分布，本身就能取到0、1，也就是说，当 $Y_i = 0, 1$ 时， $P(Y_i = 0 \text{ or } 1|\theta)$ 的全概率公式中就会有上述四种情况；而当 $Y_i > 1$ 时， $P(Y_i|\theta)$ 的全概率公式中就只有情况3和情况4，将上述事件可以等价地写为：

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

由上式和关系 $Y_i = (1 - D_i)M_i + D_i Z_i$ ，可写出更新 Z, M, D 的概率公式：

$$\begin{aligned}
P(D_i = a, M_i = b, Z_i = k | Y_i = 0, \theta) &= \begin{cases} \frac{(1-\pi_0)pP(Z_i=0)}{\pi_0(1-p)+(1-\pi_0)P(Z_i=0)}, & \text{if } k = 0, a = b = 1, \\ \frac{(1-\pi_0)(1-p)P(Z_i=0)}{\pi_0(1-p)+(1-\pi_0)P(Z_i=0)}, & \text{if } k = 0, a = 1, b = 0, \\ \frac{\pi_0(1-p)P(Z_i=k)}{\pi_0(1-p)+(1-\pi_0)P(Z_i=0)}, & \text{if } a = b = 0, k = 0, 1, \dots, \\ 0, & \text{otherwise;} \end{cases} \\
P(D_i = a, M_i = b, Z_i = k | Y_i = 1, \theta) &= \begin{cases} \frac{(1-\pi_0)pP(Z_i=1)}{\pi_0p+(1-\pi_0)P(Z_i=1)}, & \text{if } k = a = b = 1, \\ \frac{(1-\pi_0)(1-p)P(Z_i=1)}{\pi_0p+(1-\pi_0)P(Z_i=1)}, & \text{if } k = a = 1, b = 0, \\ \frac{\pi_0pP(Z_i=k)}{\pi_0p+(1-\pi_0)P(Z_i=1)}, & \text{if } a = 0, b = 1, k = 0, 1, \dots, \\ 0, & \text{otherwise;} \end{cases} \\
P(D_i = a, M_i = b, Z_i = k | Y_i = k, \theta) &= \begin{cases} 1-p, & \text{if } a = b = 0, k = 2, 3, \dots, \\ p, & \text{if } a = 0, b = 1, k = 2, 3, \dots, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{11}$$

其中 $P(Z_i = k | \theta)$ 由以下概率分布列生成：

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

更新规则如下：

- 如果 $Y_i = 0$

抛一枚正面向上概率为 $\frac{\pi_0(1-p)}{\pi_0(1-p) + (1-\pi_0)P(Z_i=0)}$ 的硬币：

- 如果正面向上， $D_i = M_i = 0$ ，按式 (12) 抽取 Z_i
- 如果反面向上， $D_i = 1, Z_i = 0$ ，再抛一枚正面向上概率为 p 的硬币，此时正面则 $M_i = 1$ ，反之为 $M_i = 0$

- 如果 $Y_i = 1$

抛一枚正面向上概率为 $\frac{\pi_0p}{\pi_0p + (1-\pi_0)P(Z_i=1)}$ 的硬币：

- 如果正面向上， $D_i = 0, M_i = 1$ ，按式 (12) 抽取 Z_i
- 如果反面向上， $D_i = 1, Z_i = 0$ ，再抛一枚正面向上概率为 p 的硬币，此时正面则 $M_i = 1$ ，反之为 $M_i = 0$

- 如果 $Y_i = k > 1$

抛一枚正面向上概率为 p 的硬币：

- 如果正面向上， $D_i = 1, M_i = 1, Z_i = k$
- 如果反面向上， $D_i = 1, M_i = 0, Z_i = k$

抽样方法

为了对上述参数进行抽样，观察各参数满条件分布式，根据式（9）（10）可以**利用Beta分布对参数 π_0, p 进行抽样**， D_i 的更新可以根据式（11）判定情形，然后利用离散分布列来抽样。

但是式（7）（8）并不是常见的分布，考虑Gibbs对其他参数更新的便捷性仍然考虑Gibbs抽样，故在对参数 $\tilde{\beta}, \phi$ 更新时通常考虑采用**M-H抽样算法**。[6][7]均提及在Gibbs抽样中混合对某部分参数M-H抽样的可行性。

1. M-H抽样算法：

Metropolis-Hastings 算法是一种构造转移核的通用方法，其基本思想和舍选抽样基本一致。

对于目标转移核，将其分解为

$$p(x, x') = q(x, x')\alpha(x, x')$$

分解式中：

- 潜在的转移核 $q(x, x')$ 作为 x 的函数是一个概率密度或概率分布，称为建议分布。

建议分布可以取各种形式，常把它取为易于产生随机数的分布。

- 评议函数 $\alpha(x, x')$ 是一个接受概率（ $0 < \alpha(x, x') \leq 1$ ），用来决定是否接受建议分布产生的潜在转移。

M-H方法的具体实施：



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。

实际计算中，为了实现上述效果，通常会产生区间 $[0, 1]$ 上均匀分布的随机数 u 与 $\alpha(x, x')$ 比较，然后根据规则进行更新，故而上述过程**实际上的一步转移**可以写为：

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

该转移形式可以对应为转移核 $p(x, x')$ 。在选定提议分布后，只剩下接受概率 $\alpha(x, x')$ 仍未确定，为了使目标 $\pi(x)$ 成为平稳分布，选择 $\alpha(\cdot, \cdot)$ 就需要一种特殊的构造，最常用的选择是：

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

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

为了**简化模型**，不妨简单取潜在的转移核 $q(x, x')$ 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

2. 引入 Pólya–Gamma 潜变量

李蒙[8]对 $\pi(\tilde{\beta}, \phi | -, Y, D, M, Z)$ 联合后验形式进行M-H抽样来更新，但**拒绝率相对较高、抽样效率低**。Qing He[1]、肖翔[2]**引入 Pólya–Gamma 潜变量，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**，施行过程见下：

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 $\kappa = a - \frac{b}{2}$ 。

将上述引理应用于式 (7)，得到：

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

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。记 $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ 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：

$$\begin{aligned} \pi(\tilde{\beta} | -, Y, D, M, Z, \omega) &\propto \pi(\tilde{\beta}) \times \prod_{i=1}^n \exp \left\{ \tilde{\kappa}_i X_i^T \tilde{\beta} - \frac{\omega_i (X_i^T \tilde{\beta})^2}{2} \right\} \\ &\propto \pi(\tilde{\beta}) \times \prod_{i=1}^n \exp \left\{ -\frac{\omega_i}{2} \left(\frac{\tilde{\kappa}_i}{\omega_i} - X_i^T \tilde{\beta} \right)^2 \right\} \\ &\propto \pi(\tilde{\beta}) \times \exp \left\{ -\frac{1}{2} (\lambda - X \tilde{\beta})^T \Omega (\lambda - X \tilde{\beta}) \right\} \end{aligned} \quad (13)$$

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

由式 (13) 可以看出 $\pi(\tilde{\beta} | -, Y, D, M, Z, \omega)$ 服从条件高斯分布. 因此由式 (13) 抽样得到 $\tilde{\beta}$ 的后验样本，公式为：

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

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

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

$\tilde{\kappa} = (\frac{(Z_1 - \phi)D_1}{2}, \frac{(Z_2 - \phi)D_2}{2}, \dots, \frac{(Z_n - \phi)D_n}{2})$, 由此即可实现对 $\tilde{\beta}$ 的抽样。

注意, 当 $D_i = 0$ 时, $PG((Z_i + \phi)D_i, 0)$ 的第一个参数将会是0, 这是不允许的。实际上, 在对 $\tilde{\beta}$ 做推断时本就应该使用 $D_i = 1$ 部分的样本, 我们在 $D_i = 0$ 时不妨直接令对应的 $\omega_i = 0$, 在后续的矩阵乘法中可以发现这样可以自动跳过对 $D_i = 0$ 样本的利用, 从而间接的实现 $D_i = 1$ 的筛选。

对于参数 ϕ , Qing He[1]、肖翔[2]均将其设置为已知参数并未在MCMC方法中估计, 但Qing He[1]建议离散参数 ϕ 可以连续化后, 使用以下方法获得:

- 使用**Metropolis - Hastings算法** (其中应用均匀先验, 从以 ϕ 的当前值为中心的零截断正态提议中抽取 ϕ 的正候选值)
- 使用**两阶段Gibbs采样** (Zhou和Carin, 2015年) [3]。

但是通常来讲负二项分布的参数 ϕ 大多时候设置先验为Gamma分布, 且方法1中的M-H算法存在问题已在上述讨论, 故我们在此处介绍方法2的**两阶段Gibbs抽样**[3]。(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)

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 $\lambda \sim \text{Gamma}(\phi, (1 - p)/p)$, 然后生成 $y|\lambda \sim \text{Pois}(\lambda)$ 来得到 $y|\phi, p \sim \text{NB}(\phi, p)$ 。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Carin (2015) [9]。

具体来说, 给定负二项模型 $y_j|\phi, p \stackrel{iid}{\sim} \text{NB}(\phi, p)$, 对于 $j = 1, \dots, n$, 先验设定为 $p \sim \text{Beta}(a_0, b_0)$, $\phi \sim \text{Gamma}(e_0, f_0)$, 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:



step1: $(p|-) \sim \text{Beta}(a_0 + \sum_j y_j, b_0 + n\phi)$

step2: $(l_j|-) \sim \text{CRT}(y_j, \phi)$

step3: $(\phi|-) \sim \text{Gamma}(e_0 + \sum_j l_j, f_0 - n \log(1 - p))$

在上述展示中, “ $|-$ ”表示给定数据和所有其余参数的条件分布。对 p 进行采样的第一步源自贝塔 - 二项共轭性。最后两步完成参数 ϕ 的采样, 这也涉及辅助变量 l_j 。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 $b_m \sim \text{Bernoulli}(\phi/(m - 1 + \phi))$, 我们记 $(l_j|-) \sim \text{CRT}(y_j, \phi)$ 。

现在，给定所有的 l_j ， ϕ 可以通过第三个方程进行采样，该方程是通过伽马 - 泊松共轭性得到的。

在我们的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 $n_j, j = 1, 2, \dots, h$ ，然后利用这些数据进行更新：

$$\begin{aligned}
 \text{step1: } & (l_j | -, Y, D^{(t-1)}, M^{(t-1)}, Z^{(t-1)}) \sim CRT(Y_{n_j}, \phi^{(t-1)}) \\
 \text{step2: } & \gamma_{n_j}^{(t-1)} = \frac{\mu_{n_j}^{(t-1)}}{\mu_{n_j}^{(t-1)} + \phi^{(t-1)}} \\
 \text{step3: } & (\phi^{(t+1)} | -, Y, D^{(t-1)}, M^{(t-1)}, Z^{(t-1)}) \sim Gamma(e_0 + \sum_j^h l_j, f_0 - \sum_{j=1}^h \log(1 - \gamma_{n_j}^{(t-1)}))
 \end{aligned} \tag{15}$$

在迭代结束后，可以考虑将估计参数 $\hat{\phi}$ 进行取整操作。

伪代码：

至此，上述参数均可在Gibbs抽样中进行每个维度参数的更新，伪代码形式的流程如下：

Algorithm 1: Gibbs 抽样流程

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

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

1 **Initialize:**

2 **目标参数初始化:** $\tilde{\beta}^{(0)}, \phi^{(0)}, \pi_0^{(0)}, p^{(0)}$

3 **隐含数据初始化:**

4 $D_i = 1$ if $Y_i \geq 1$ 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

5 $M_i = 1 \sim \text{Bernoulli}(0.5)$

▷ 随机由 Bernoulli 分布分配

6 $Z_i = Y_i$ if $Y_i \geq 1$ 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

7 **先验参数初始化:** $\mu_{\tilde{\beta}}, \sigma_{\tilde{\beta}}^2, e_0, f_0$

8 $t \leftarrow 1, \text{Iteration}, \text{Burn-in}$

▷ 设置迭代次数与 Burn-in 长度

9 **for** $t = 1, 2, \dots, \text{Iteration}$ **do**

10 (a) **更新** $\gamma_i^{(t-1)}$

▷ 需要参数: $\tilde{\beta}^{(t-1)}$

11 $\gamma_i^{(t)} = \frac{\exp(X_i^T \tilde{\beta}^{(t)})}{1 + \exp(X_i^T \tilde{\beta}^{(t)})}$

12 (b) **更新** $\pi_0^{(t)}$

▷ 利用 R 软件从 Beta 分布抽样得到

13 $\pi_0^{(t)} \sim \text{Beta}(n + 1 - \sum_{i=1}^n D_i^{(t)}, 1 + \sum_{i=1}^n D_i^{(t)})$

14 (c) **更新** $p^{(t)}$

▷ 利用 R 软件从 Beta 分布抽样得到

15 $p^{(t)} \sim \text{Beta}(\sum_{i=1}^n M_i(1 - D_i^{(t)}) + 1, \sum_{i=1}^n (1 - M_i)(1 - D_i^{(t)}) + 1)$

16 (d) **更新隐含样本** $D^{(t)}, M^{(t)}, Z^{(t)}$

▷ 利用 $P(D_i = a, M_i = b, Z_i = k | Y_i = c, \theta^{(t-1)})$ 条件概率抽样得到

18 $D_i^{(t)}, M_i^{(t)}, Z_i^{(t)} \sim P(D_i = a, M_i = b, Z_i = k | Y_i = c, \theta^{(t-1)}), \ i = 1, 2, \dots, n$

19 (e) **更新** $\tilde{\beta}^{(t)}$

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

21 **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

22 $\omega_i \sim \text{PG}((Z_i + \phi^{(t-1)})D_i^{(t-1)}, 0), \ i = 1, 2, \dots, n$, 记 $\Omega = \text{diag}(\omega_1, \omega_2, \dots, \omega_n)$

▷ 生成隐变量

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

24 $\tilde{\kappa} = (\frac{(Z_1 - \phi)D_1^{(t-1)}}{2}, \frac{(Z_2 - \phi)D_2^{(t-1)}}{2}, \dots, \frac{(Z_n - \phi)D_n^{(t-1)}}{2})$

25 $\tilde{\beta}^{(t)} | -, Y, D^{(t-1)}, M^{(t-1)}, Z^{(t-1)}, \omega \sim N(\tilde{M}_\omega, \tilde{H}_\omega)$

▷ 正态抽样

26 (f) **更新** $\phi^{(t)}$

27 **方法 1:** M-H 抽样法 -> $\phi^{(t)}$

28 **方法 2:** 通过 CRT 的两阶段 Gibbs 抽样

▷ 利用公式 (15) 抽样得到

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

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

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

32 **end**

33 保存去掉 Burn-in 后的数据。

实验结果及分析

数据分析与可视化

针对作业所给出的数据, 我们作以下可视化处理:

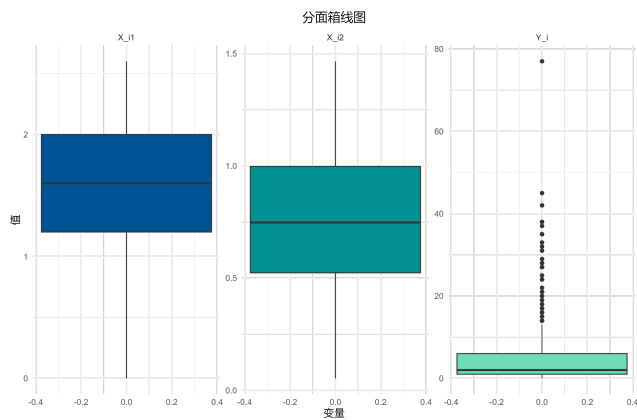


Figure: 箱线图

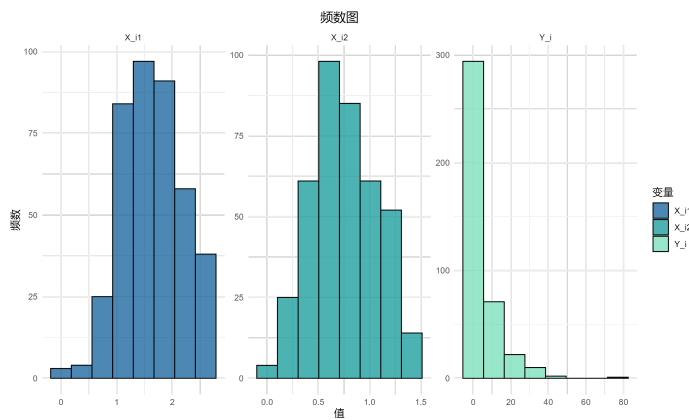


Figure: 频数图

分面箱线图和频数图可以反映出数据的**离散分布情况**。

从图中可以看出：

- 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**。

相关性图

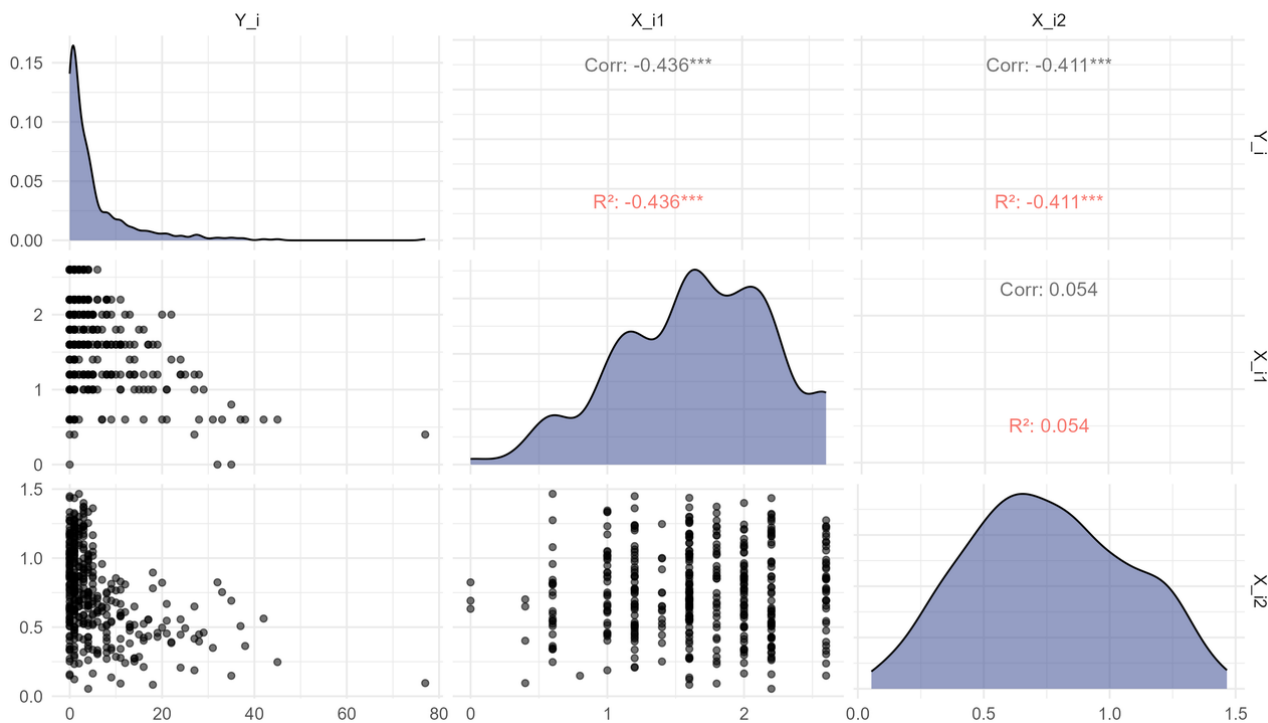


Figure: 相关性图

相关性图对角线上的**密度图**提供了每个变量的分布概况，可以看出： Y_i 和 X_{i1} 的分布的偏度确实如前所述呈现偏态，而 X_{i2} 的分布状如正态，有轻微右偏；

下三角的**散点图**则提供了每两个变量之间的具体关系，可以看出： Y_i 的异常值所对应的 X_{i1} 和 X_{i2} 数值均偏小， X_{i2} 在较小值一侧对应的 X_{i1} 较离散，在较大一侧对应的 X_{i1} 较集中；

上三角展示了每两个变量之间的**相关系数**， Y_i 与 X_{i1} 及 X_{i2} 均呈现出负相关关系， X_{i1} 与 X_{i2} 之间为正相关关系。

在正式对给出未知参数的数据集 `Data_0-1膨胀负二项回归.xlsx` 使用MCMC方法估计前，我们通过自行设定的真实参数和模拟生成测试数据集检验方法估计的可靠性。

在下面的实验过程中，我们将逐步完成以下的检验实验：

生成数据

通过上述问题分析已经知道， Y_i 与 X_{i1} 及 X_{i2} 均呈现出**负相关关系**，故而在生成数据时同样考虑将 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，需要设置较高的 β_0 。

为了进一步提高对测试数据的估计精度，不妨直接采用 `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

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

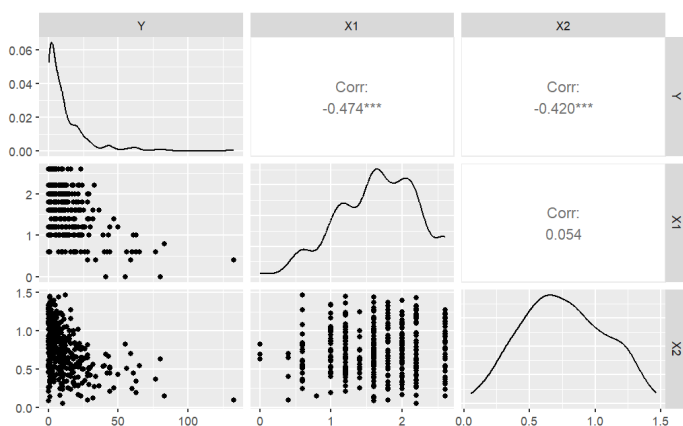
```
1 # 生成自变量 X1, X2 (假设是正态分布)
2 n <- 400
3 X1 <- test_data$X_i1
4 X2 <- test_data$X_i2
5 X <- cbind(1, X1, X2) # 包括常数项
```

```

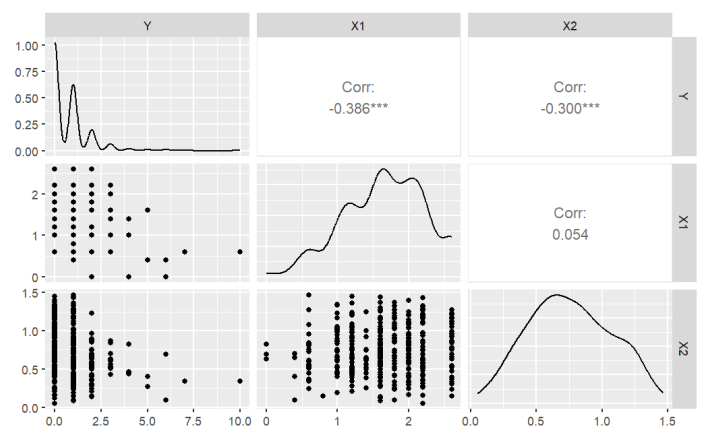
6
7 # 给定回归系数和生成参数
8 beta_true <- c(5.5, -1, -2)
9 phi_true <- 4
10 p_true <- 0.6
11 pi0_true <- 0.2
12
13 # 计算mu_i
14 mu <- exp(X %*% beta_true)
15 gamma <- mu/(mu+phi_true)
16 # 根据mu和phi生成Z (负二项分布)
17 Z <- rbinom(n, size = phi_true, mu = mu)
18 M <- rbinom(n, 1, p_true) # M_i
19 # 根据D和Z生成Y
20 D <- rbinom(n, 1, 1 - pi0_true) # D_i
21
22 Y <- (1 - D) * M + D * Z # 生成观测数据Y
23 data <- data.frame(Y, X1, X2)

```

生成数据的大致分布情况如下（下图左），可以看到生成数据集和 `Data_0-1膨胀负二项回归.xlsx` 数据之间的相关情况与分布均类似，可以作为一个较好的评估方法有效性的**测试数据集**：



图左：合适的测试数据



图右：不合适的测试数据

另外，当我们生成数据时，**需要控制 μ_i 不能均一致过小**，否则负二项分布的数据均值 $\mathbb{E}(Z_i) = \mu_i$ 会导致实际的 Z_i 本身很小，并有相当大的概率在0-1之间（上右图），此时模型中0-1值内二项分布和负二项分布的混合程度很大，后续模型更新隐含数据 D, M, Z 的效果将非常差，从而将不属于负二项分布/二项分布的数据分给对应分布的Gibbs抽样步骤中进行参数估计，导致较大偏差。

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提及的Gibbs抽样方法，检测估计效果的好坏。

(1) 更新隐含数据 D, M, Z

初始化:

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

```
1 D[1, ] <- ifelse(Y >= 1, 1, rbinom(n, 1, 0.5)) # 初始化 D
2 M[1, ] <- rbinom(n, 1, 0.5) # 初始化 M
3 Z[1, ] <- ifelse(Y > 1, Y, rbinom(n, 1, 0.5)) # 初始化 Z
```

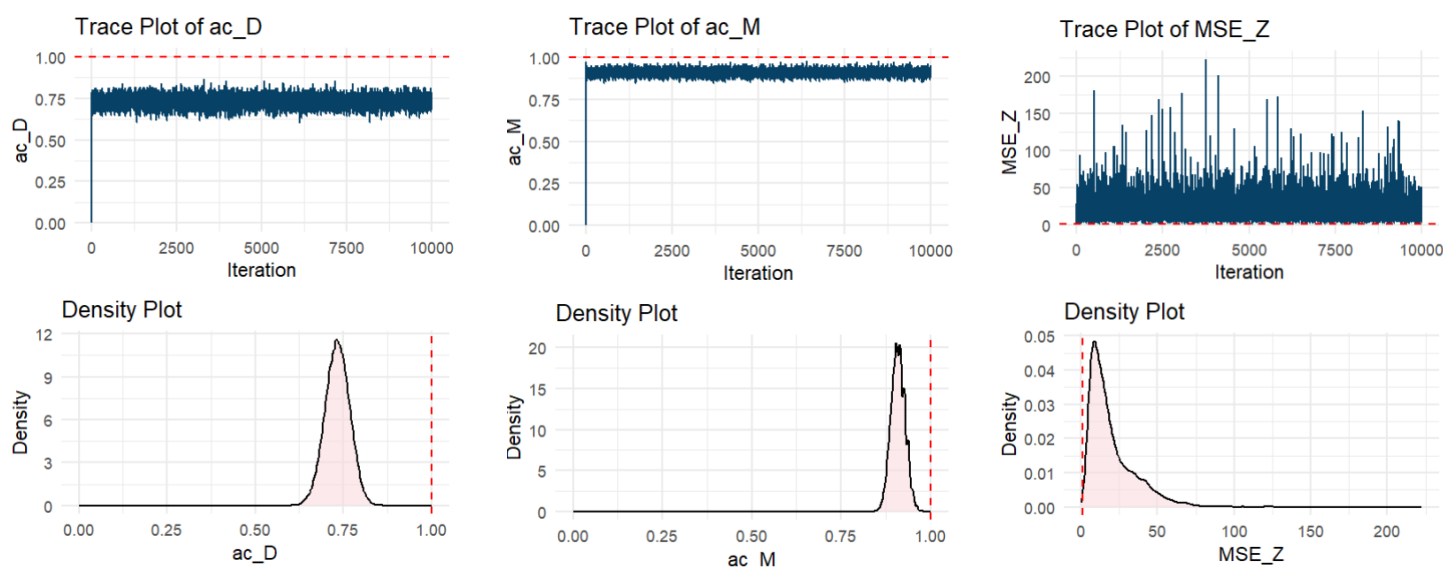
在进行**10000次**更新过程中，我们将每次更新的 $D^{(t)}, M^{(t)}, Z^{(t)}$ 与完全数据进行比较。由于当 $Y > 1$ 时一定属于负二项分布，加入这一部分的准确率判断没有什么意义，于是我们只关注 $Y = 0, 1$ 时，上述隐含数据的估计效果。

由于此后几种待估计的参数抽样方法均是利用 $D^{(t)}$ 先将数据分为来自负二项分布和二项分布的样本，然后各自更新两类样本对应分布的参数，也就是说我们实际关心的是以下几种问题：

- 当 $D_i^{(t)}$ 把样本 Y_i 划分为负二项分布时， $Z_i^{(t)}$ 和真实的 Z_i 差距为多少？考虑用MSE来刻画所有被划分为负二项分布的样本与真实数据的差异。
- 当 $D_i^{(t)}$ 把样本 Y_i 划分为二项分布时， $M_i^{(t)}$ 和真实的 M_i 差距为多少？由于 M 和 D 均为取值在0-1，考虑用**准确率**来刻画所有被分为二项分布与真实数据的差异。

因此，在Gibbs抽样的过程中，我们可以额外监控每一步迭代时生成的隐含数据 $D^{(t)}, M^{(t)}, Z^{(t)}$ 和真实值的差距，如果这种差距足够小，那么后续更新各个分布的参数时就可以更好的利用数据。

运行结果:



平均意义下:

- 隐含数据 $D^{(t)}$ 更新的准确率为**73.27%**
- 隐含数据 $M^{(t)}$ 更新的准确率为**90.94%**
- 隐含数据 $Z^{(t)}$ 更新的**MSE**为**20.24106**

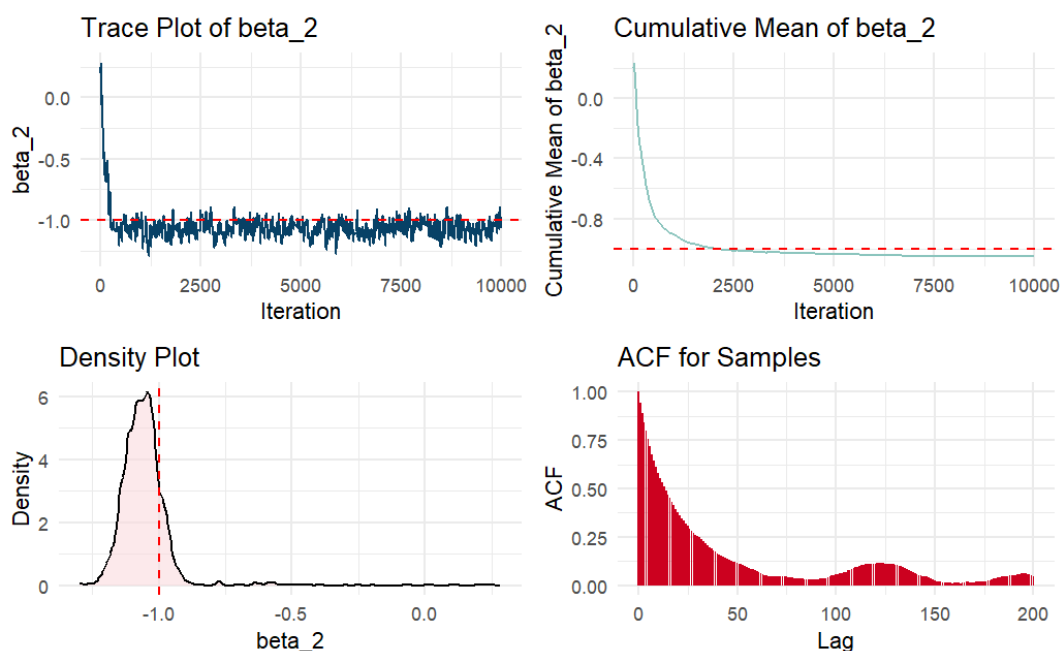
通过分布图可以观察到，对 $D^{(t)}$, $M^{(t)}$ 更新的**分布较狭长、准确率较稳定**，但对 $Z^{(t)}$ 的估计**有相对较大的偏差**。换句话说，在该数据因为未知原因仅被简单区分为0、1时，背后实际的计数并不那么容易被估计。

(2) 更新 $\tilde{\beta}$

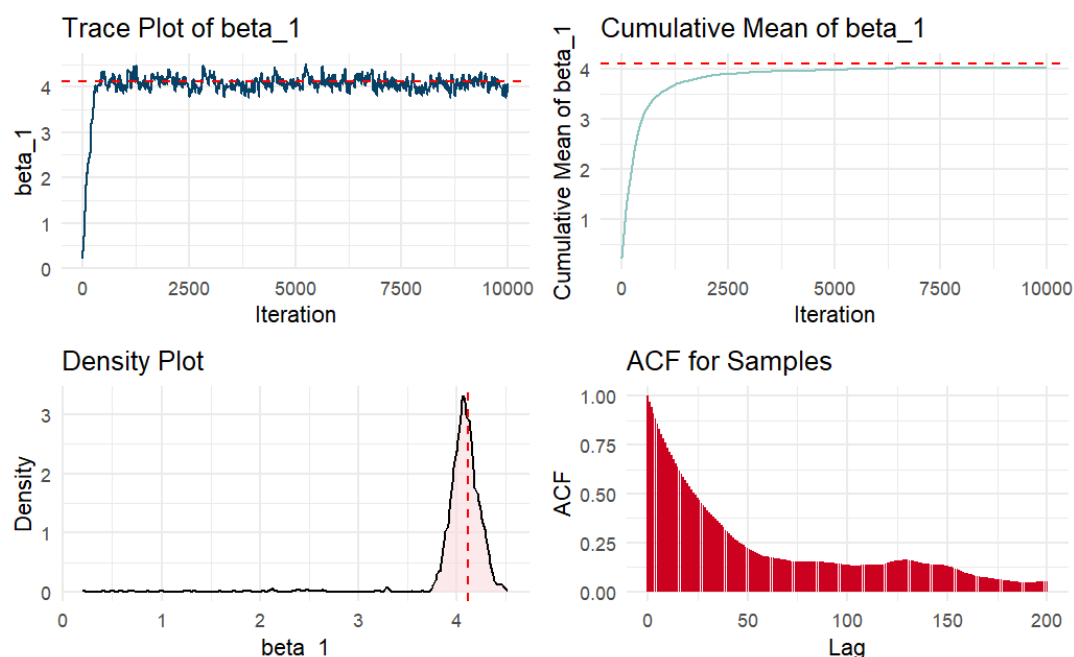
iterations	burn_in	$\tilde{\beta}^{(0)}$	$E(\beta_{prior})$	$SE(\beta_{prior})$
10000	4000	(0.2,0.2,0.2)	(4, -1, -2)	(1, 1, 1)

方法1：M-H抽样法

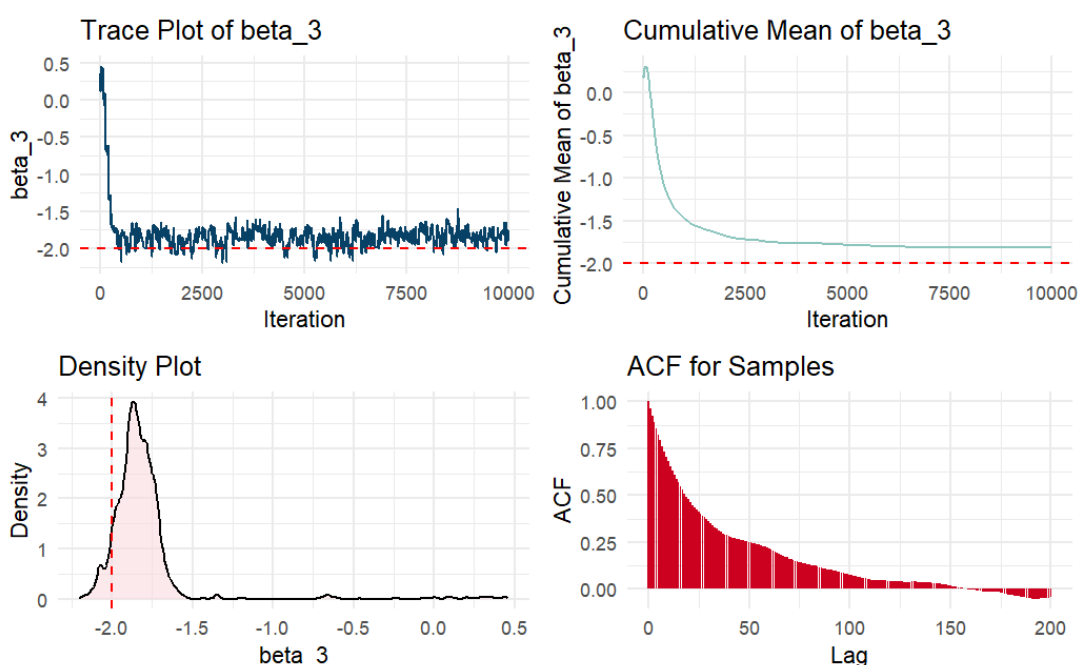
Posterior mean of beta	4.09535	-1.065358	-1.869007
true beta	4.113706	-1	-2



$\tilde{\beta}_0$ 的抽样情况



$\tilde{\beta}_1$ 的抽样情况



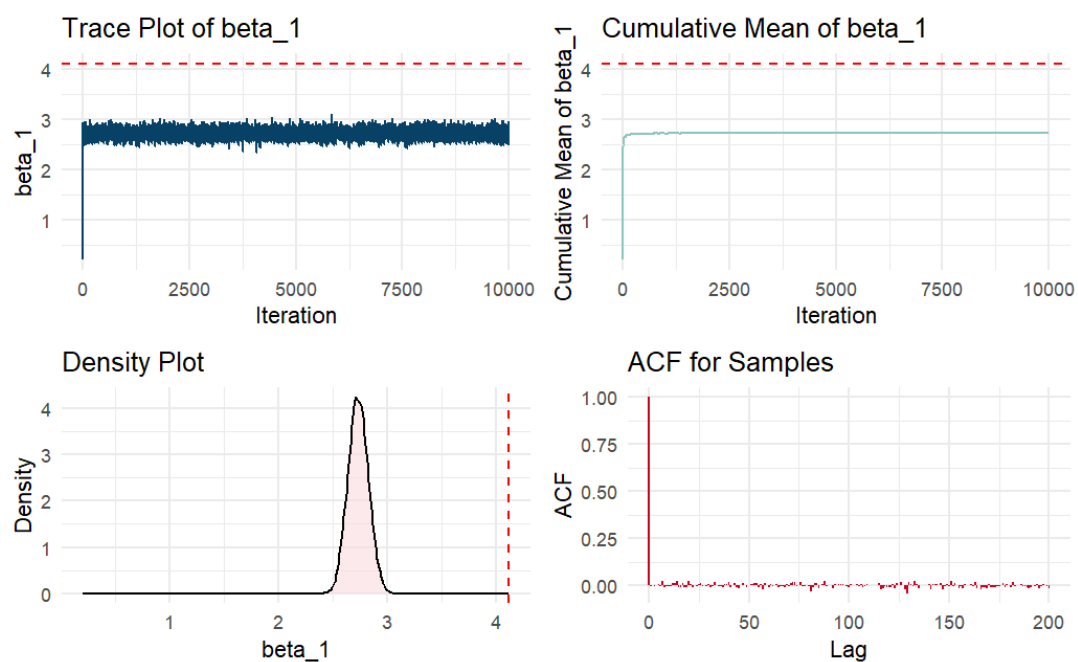
$\tilde{\beta}_2$ 的抽样情况

当迭代次数超过一万次后，三个参数的均值线已经趋于平稳，且较好地收敛到了真实值。通过图像分析可得，设定burn-in期为5000是合理的。此外，由于 Metropolis-Hastings 算法的高拒绝率，导致自相关系数较高。因此，考虑采用 **Thinning** 策略来减少自相关对结果的影响。

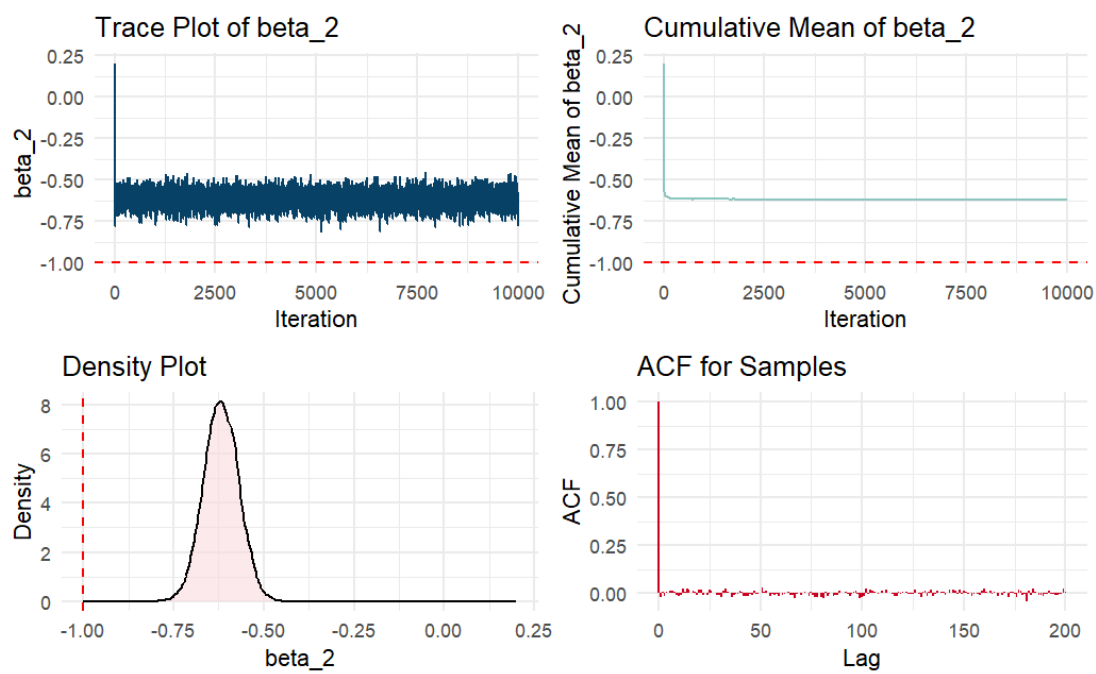
进一步分析发现，三个参数的分布基本呈现较狭长的正态分布，表明估计效果良好。

方法2: Pólya–Gamma隐变量抽样法

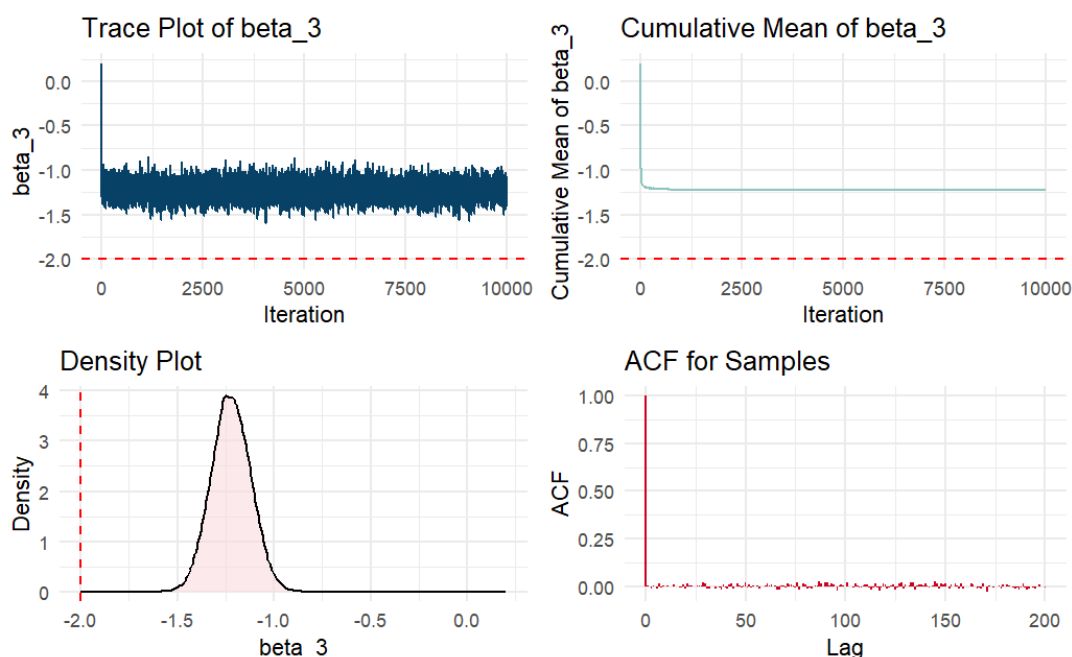
Posterior mean of beta	2.728543	-0.6184657	-1.223236
true beta	4.113706	-1	-2



$\tilde{\beta}_0$ 的抽样情况



$\tilde{\beta}_1$ 的抽样情况



$\tilde{\beta}_2$ 的抽样情况

在实际操作中发现，尽管该方法抽样时，每一步迭代都会发生转移，因此不会面临MH算法中常见的高拒绝率问题，但是当样本量较大时**矩阵求逆的速度会较大影响迭代速度**，该方法的权衡思想是通过增加**计算的复杂度**来换取**时间复杂度**的降低。

一个较为严重的问题是，该方法不知为何表现出稳定的极度有偏的现象。若上述理论推导无误，则原文可能省略了在操作过程中的一些关键细节。

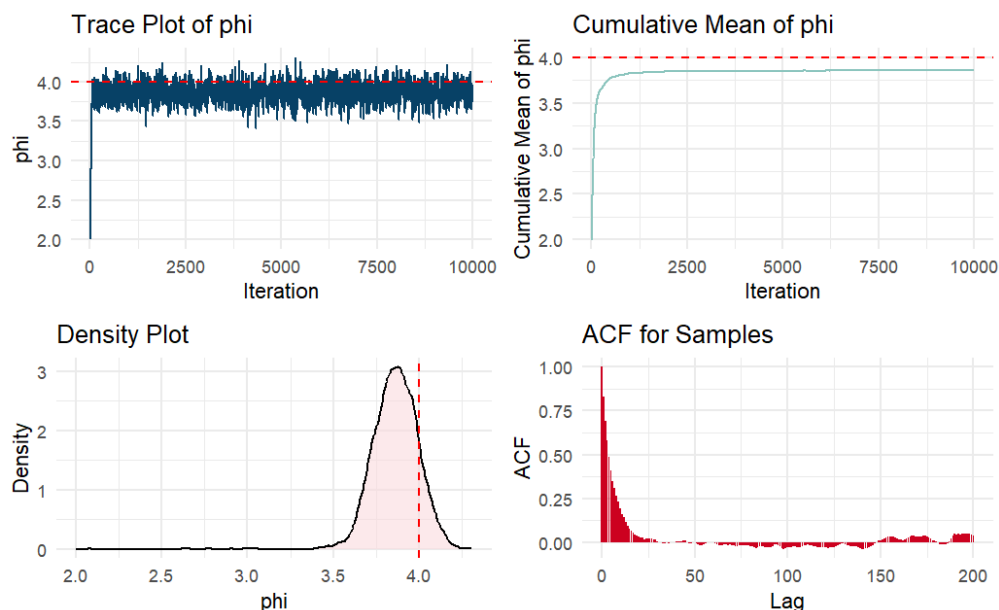
综上所述，我们决定在后续的Gibbs抽样中，继续选择使用MH算法对 $\tilde{\beta}$ 进行抽样。

(3) 更新 ϕ

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

方法1：M-H抽样法

可以看到，当迭代次数到**5000**时，均值已经基本收敛，burn-in可考虑为2000~4000。在尝试多次后发现，估计是否能收敛于真实值主要依赖生成数据的特性。在某些情况下生成的数据，即便M-H抽样算法尽力进行迭代，估计结果也仅能

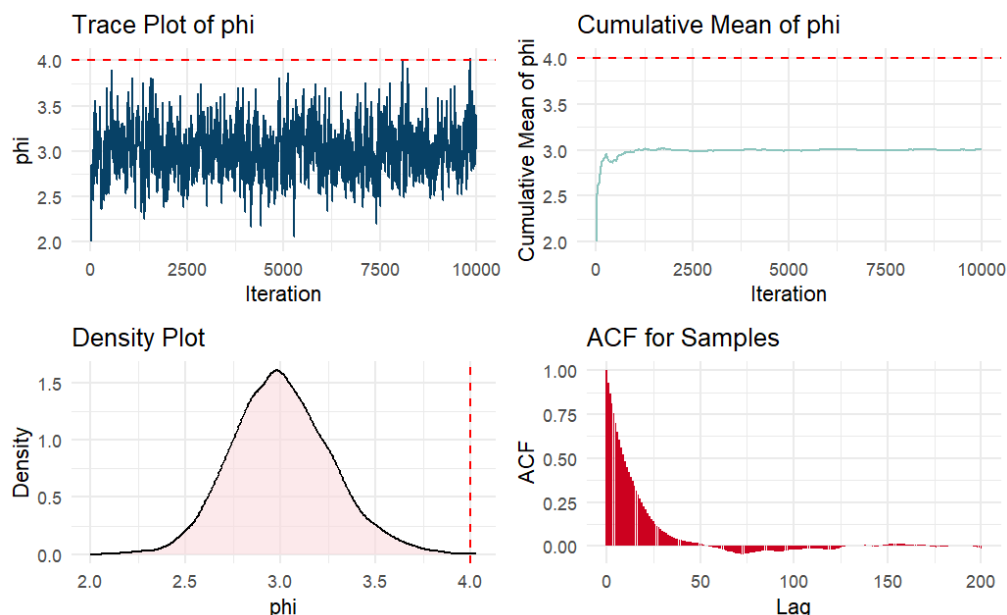


达到左图所示的精度水平，无法进一步提高。

此外，该方法下同样有高拒绝率导致的**难收敛、高自相关性问题**。此时单独更新 ϕ 还未暴露该问题的显著性，后续完整的Gibbs抽样中可以看到明显的影响。

估计后验分布均值：3.864

方法2：通过CRT的两阶段Gibbs抽样生成 ϕ



该方法依然表现出极度有偏，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。

估计后验分布均值：3.005026

综上，我们仍然选择MH算法用于后续的Gibbs抽样中对 ϕ 的抽样。

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	$= \begin{cases} 1 & Y_i \geq 1 \\ \sim b(0.5) & else \end{cases}$	$\sim b(0.5)$	$= \begin{cases} Y_i & Y_i \geq 1 \\ \sim b(0.5) & else \end{cases}$
-------------------------	---	-----	-----	--	---------------	--

超参数设置：

（1）先验分布设置：

- $E(\beta_{prior})$: 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 $N(x, \sigma_0^2)$

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

实验设置与估计结果：

迭代次数：50000

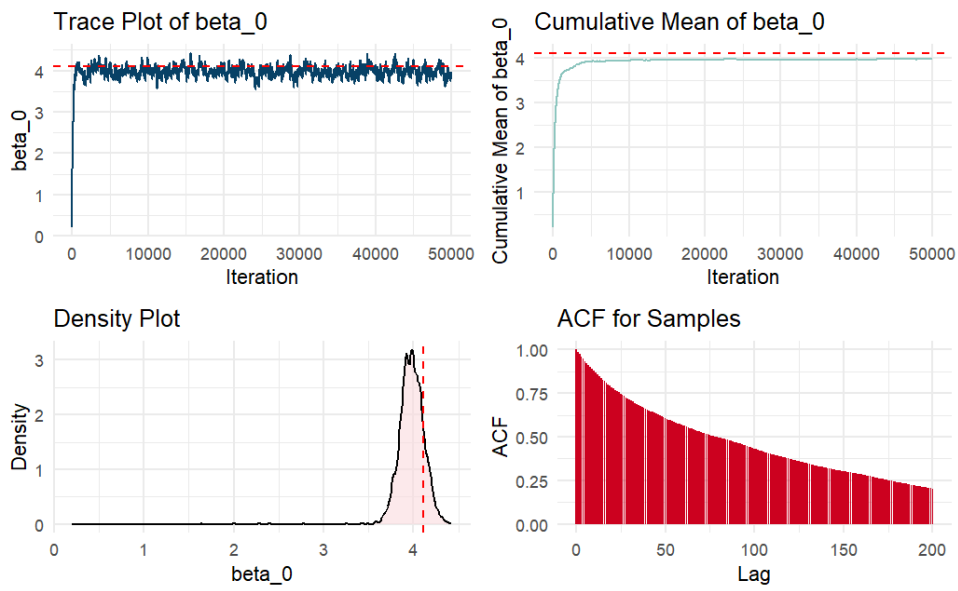
burn_in：10000

获得估计如下：

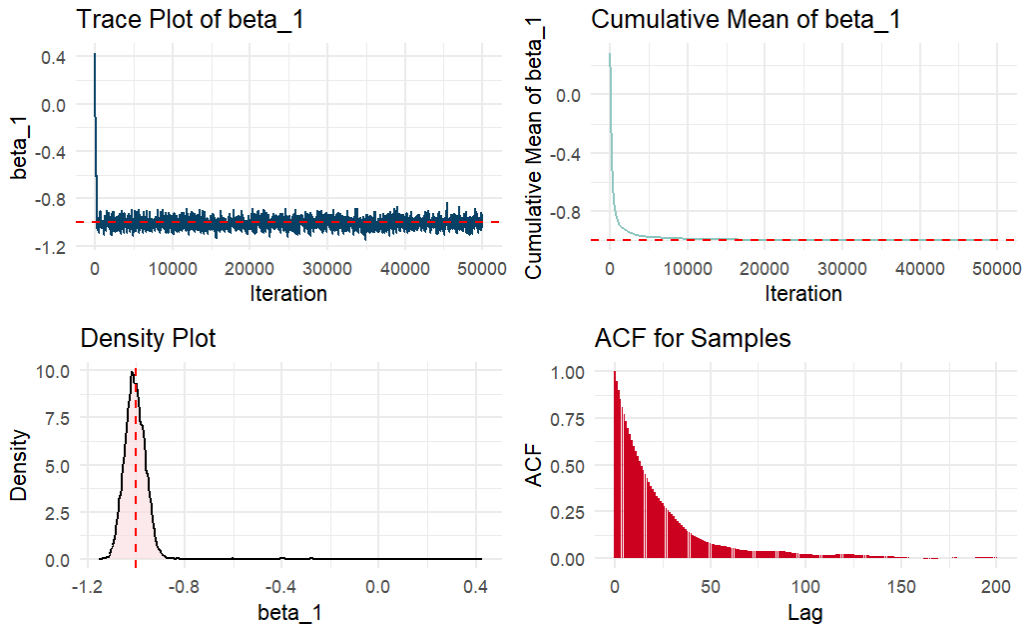
参数	β_0	β_1	β_2	ϕ	p	π_0
真实值	5.5	-1	-2	4	0.6	0.2
Gibbs抽样估计后验分布均值：	5.499993	-1.005951	-2.054578	4.572603	0.6253828	0.2075729

我们画出各参数在Gibbs抽样中迭代时的记录，进一步分析：

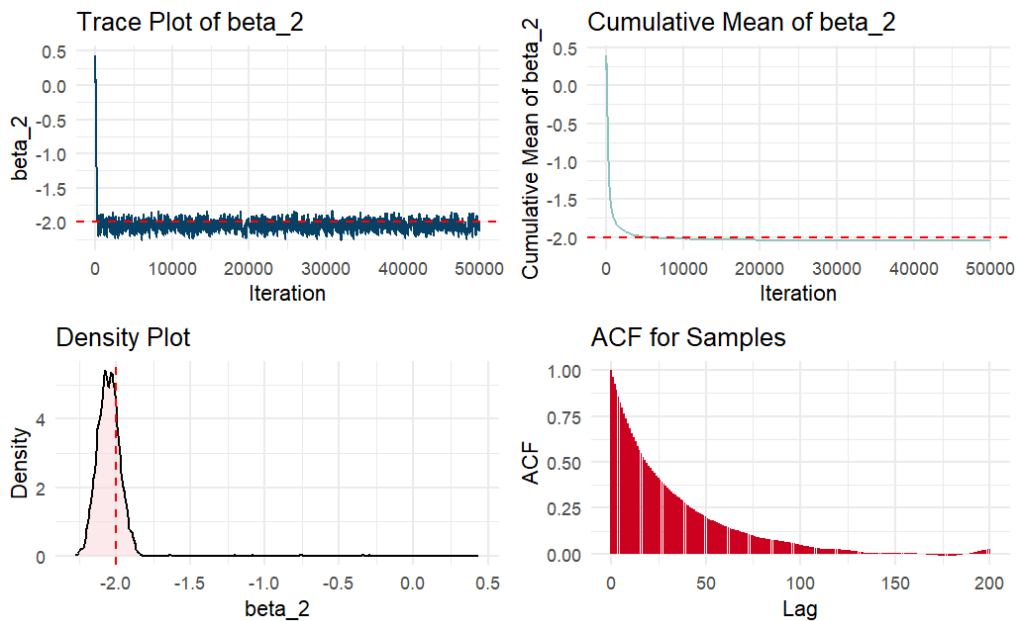
1. 回归系数 $\vec{\beta}$ ：



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



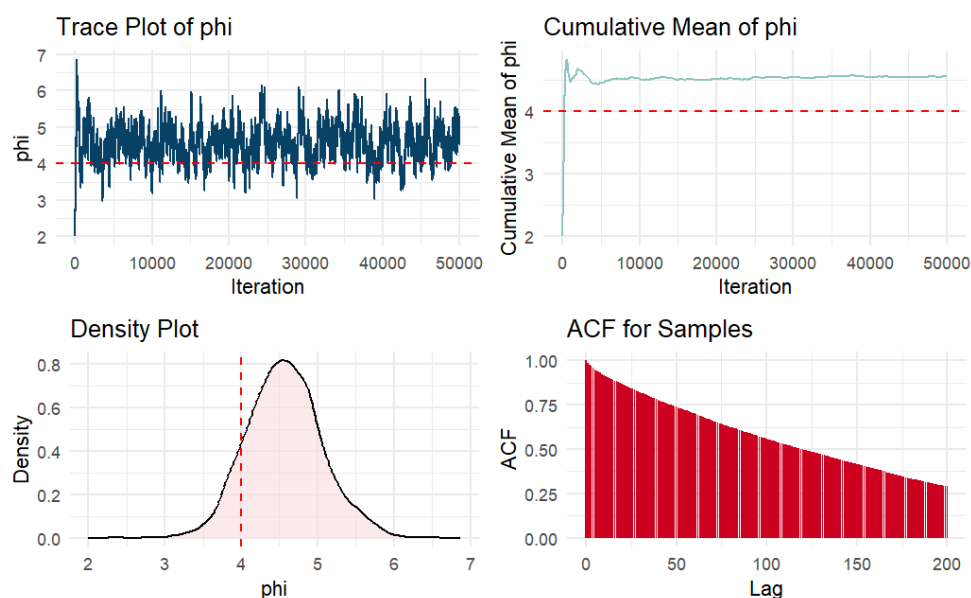
$$\beta_1$$



三个回归系数的估计基本已收敛到真实参数。

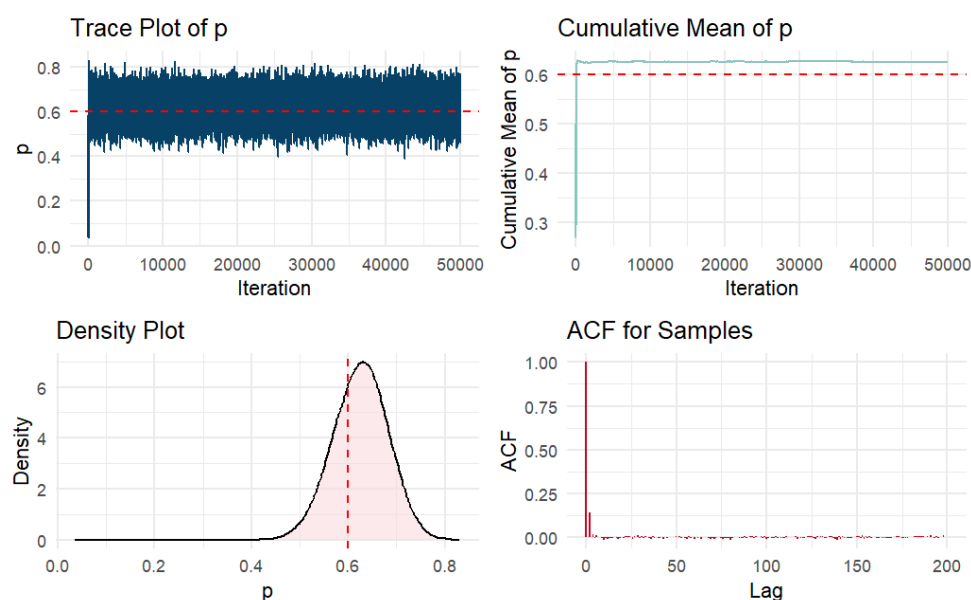
然而存在一个奇怪的现象：通过 $\tilde{\beta}_0 = \beta_0 - \log(\phi)$ 转换后得到的 β_0 估计是几乎无偏的。但由于后续会发现 ϕ 的估计存在偏差，导致 $\tilde{\beta}_0 = \beta_0 - \log(\phi)$ 仍然存在偏倚。

2. 过离散参数 ϕ : $Z_i \sim NB(\mu_i, \phi)$



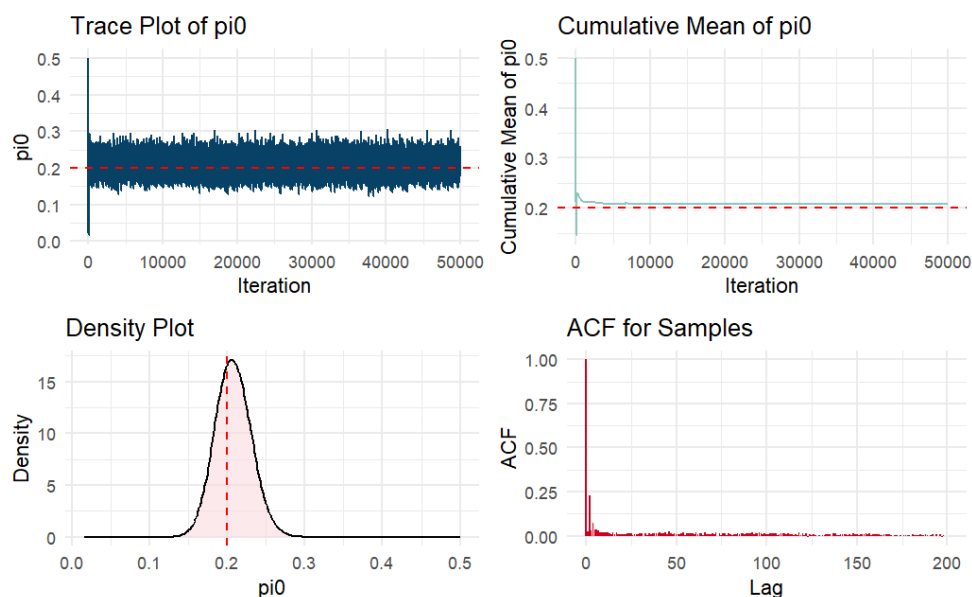
负二项分布的过离散参数 ϕ 出现了较大的偏移，从轨迹图来看， ϕ 收敛情况较差，但是累积均值图已经基本收敛。猜测更多的迭代次数也很难再有精度的改善。在当前迭代过程中， ϕ 的自相关性在所有参数中最为严重。

3. 二项分布参数 p : $M_i \sim B(1, p)$



轨迹图显示出良好的收敛态势，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。

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有显著的优势。

未知参数数据估计：

这里的基础设置和上面自设参数的数据集Gibbs抽样中基本一样，不再赘述，只额外修改了先验分布的超参数：

$E(\beta_{prior})$	$SE(\beta_{prior})$	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？

我们的解释是，通过预先进行的估计结果大致猜测真实参数的范围来设置的，这有助于模型更快的收敛结果。

实验设置与估计结果：

迭代次数：**20万次**

burn-in：**10万次**

获得估计如下：

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

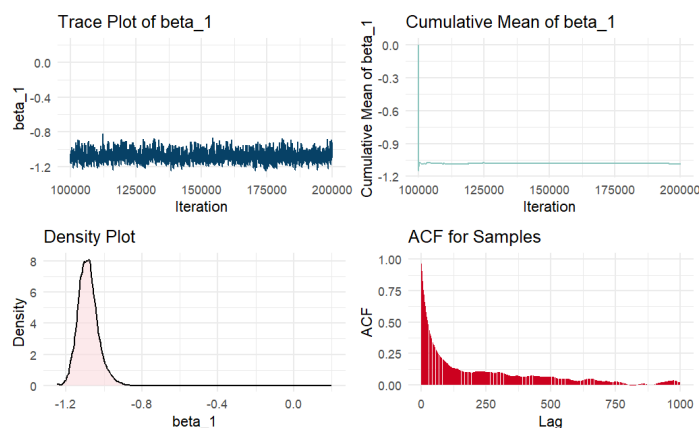
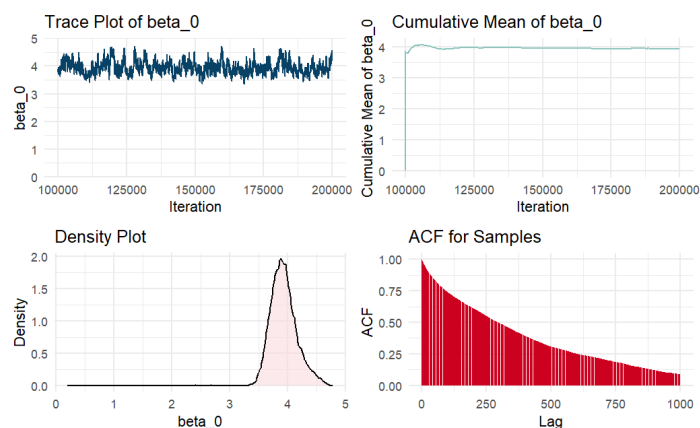
简单解释：第一行是**Gibbs抽样实际估计效果**，第二行是考虑到人为设置参数应该比较整，顺手做一个**猜测**，此外在下面的分析中可以看到 ϕ 的估计很难收敛，可能会导致其余部分均出现偏差，当迭代次数足够大时就会得到下面的结果。但是当迭代次数只有5万次时其实非常接近表中**猜测真实参数**，说不定是因为 ϕ 估歪了导致长远来看把其他参数都带歪了呢……

我们画出各参数在Gibbs抽样中迭代时的记录，进一步分析：

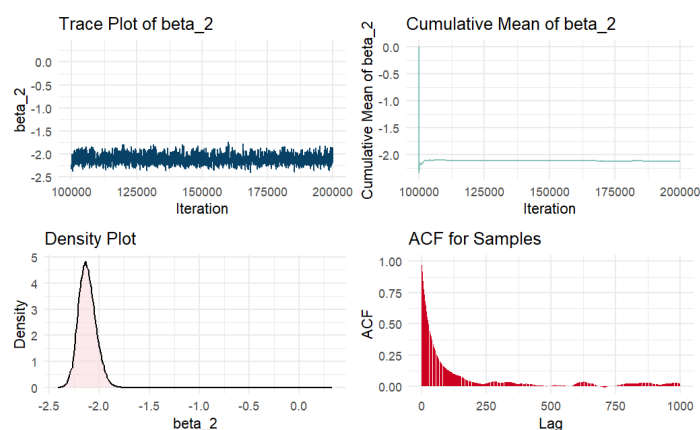
1. 回归参数：

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

$$\beta_1$$

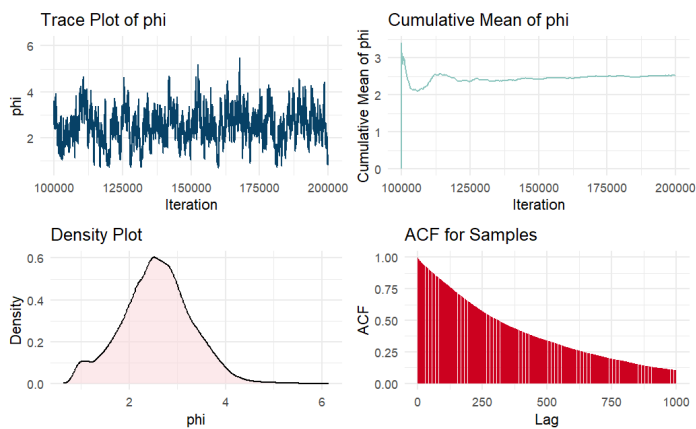


$$\beta_2$$



基本的结论是和前面的采用自己生成的数据得到的结果相似。

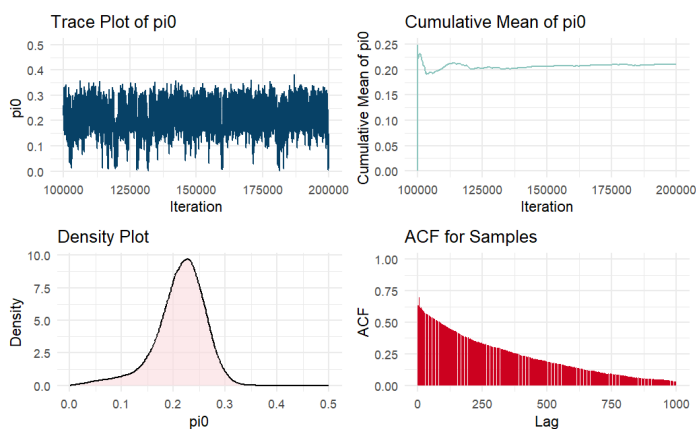
2. 负二项分布参数 ϕ ：



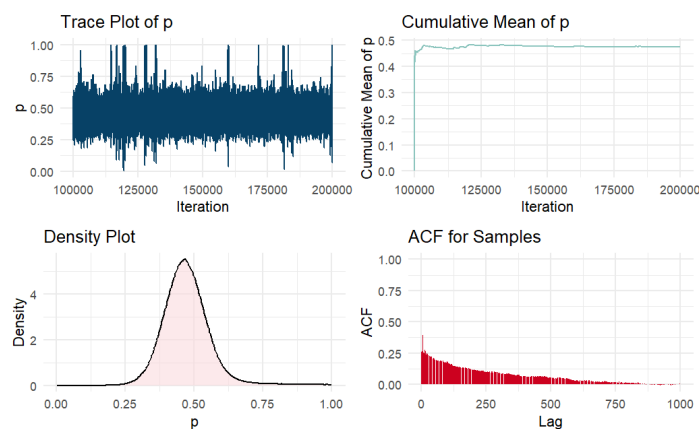
正如最开始提到，所有参数中只用 ϕ 不容易收敛，轨迹图并不理想情况的密集。

3. 二项分布参数：

π_0



p



基本较好收敛。

参考文献及资料

1. Qing He, Hsin-Hsiung Huang. A framework of zero-inflated Bayesian negative binomial regression models for spatiotemporal data. *Journal of Statistical Planning and Inference*, ISSN: 0378-3758, 2024, 229(106098).
2. 马巧玲, 肖翔. 0—1膨胀负二项回归模型在COVID—19疫情分析中的应用[J]. *上海工程技术大学学报*, 2022, 36(2): 212-217.
3. Zhao L, Wu W, Feng D, Jiang H, Nguyen X. Bayesian Analysis of RNA-Seq Data Using a Family of Negative Binomial Models. *Bayesian Anal*, 2018, 13(2): 411-436.
4. Tang, Y., Liu, W., & Xu, A. (2017). Statistical inference for zero-and-one-inflated poisson models. *Statistical Theory and Related Fields*, 1(2), 216–226.

5. Nicholas G. Polson, James G. Scott, Jesse Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables[J]. Journal of the American Statistical Association, 2013, 108(504): 1339 — 1349.
6. Mohamed G. M. Ghazal, Mustafa M. Hasaballah, Rashad M. EL-Sagheer, et al. Bayesian Analysis Using Joint Progressive Type-II Censoring Scheme. Symmetry, 2023, 15(10): 1884.
7. Fengshi Zhang, Wenhao Gui. Parameter and Reliability Inferences of Inverted Exponentiated Half-Logistic Distribution under the Progressive First-Failure Censoring. Mathematics, 2020, 8(5):708.
8. 李蒙. 0—1膨胀负二项模型及其统计分析[D]. 上海：华东师范大学, 2018.
9. Zhou Mingyuan, Carin Lawrence. Negative Binomial Process Count and Mixture Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 37(10): 1109.