arXiv:2301.01852v1 [stat.ME] 4 Jan 2023

具有序列相关误差的删失回归：贝叶斯方法

Rodney Sousa* Isabel Pereira† Maria Eduarda Silva‡
Brendan McCabe§

摘要

在许多环境研究和社会研究中，估计误差存在自相关的删失线性回归模型是一个常见问题。本文提出一种贝叶斯方法，用于估计具有 AR(p) 误差的删失回归模型。本文开发的算法采用带数据增广的吉布斯采样器（GDA），在每次迭代中对模型参数和潜变量进行抽样。通过从相应的条件分布中多次抽取潜变量来实现数据增广。适当的变量变换使得能够得到完整似然。模拟研究表明，即使在删失观测占比很高的情况下，所提出的方法仍能获得精度较高的估计值。

本文还通过云底高度的真实数据进一步阐述了该方法，包括对删失时间序列数据进行模型检验和模型选择。

关键词：删失数据、线性回归、自相关、贝叶斯分析、吉布斯采样器、数据增广

1 引言

当对观测数据设定明确的界限时，就会出现删失观测；这种情况见于多个领域，包括环境监测、经济学、医学和社会科学。删失可能是由测量设备的局限所致，例如空气污染或水中矿物质浓度的检测限（Hopke等人，2001）。在经济学中，当施加约束或法规时也会出现删失，例如国际贸易中的观测数据，出口和进口会受到贸易壁垒的限制（Zangari 和 Tsurumi，1996）。

*通讯作者：rodney@ua.pt，葡萄牙阿威罗大学和 CIDMA
†isabel.pereira@ua.pt，阿威罗大学数学系，葡萄牙，CIDMA

‡mesilva@fep.up.pt，波尔图大学经济学院，葡萄牙，LIADD-INESC§Brendan.Mccabe@liverpool.ac.uk，利物浦大学管理学院，查塔姆大楼，查塔姆街，邮编 L69 7ZH

自 Buckley 和 James（1979）的研究以来，关于删失响应回归分析的大量文献不断涌现。除删失之外，数据往往还呈现序列相关，因此人们采用了动态删失模型。

在时间序列回归的背景下，已有多位作者研究了删失问题。Zeger 和 Brookmeyer（1986）提出了第一种用于估计具有自相关误差的删失回归的方法，并给出了该模型的精确似然函数。该似然函数基于维度可变的数据块构建。由于数据块大小通常会随删失率的升高而增大，最大似然估计很快便会在数值计算上变得难以处理。鉴于这一问题，作者建议采用基于伪似然的近似方法。Park 等人（2007）引入了一种插补法，用于从删失时间序列中估计 ARMA 模型。根据观测数据和删失信息给出的条件分布，通过随机模拟生成的值对可能被删失的值进行插补。由此得到的时间序列被视为完整数据，可以使用常规时间序列方法进行分析。Mohammad（2014）提出了一种准 EM 算法，用于拟合存在删失的 ARMA 模型，其特点是将缺失数据视为删失的一种特殊情况。Schumacher 等（2017）建议使用 EM 算法的随机近似（SAEM），该方法基于具有 AR(p) 误差的线性回归模型的无条件似然函数。这些作者通过模拟表明，即使删失值的比例很高（$\approx 40\%$），该方法仍能得到相合估计。Houseman 和 Virji （ 2017 ）提出了一种贝叶斯方法，用于处理存在左删失的暴露时间序列数据；该方法采用基于样条的方法对自相关进行建模，以考虑非平稳自相关。Wang 和 Chan（2018）提出了一种基于方程组的准似然法，并根据模拟残差进行了模型检验（Gourieroux 等，1987）。

基于删失观测估计存在自相关误差的回归模型这一问题，也曾在贝叶斯框架下得到研究。Zangari 和 Tsurumi（1996）针对存在 AR(1) 误差的删失回归模型，考察了三种贝叶斯程序。作者以 Zeger 和 Brookmeyer（1986）的研究为基础，采用拉普拉斯近似、结合数据增广的吉布斯采样器以及求积数值积分程序，推导出模型参数的后验密度。然而，作者发现，在删失比例适中（10-15%）且扰动项存在较强相关性的情况下，采用数据增广算法的吉布斯采样器无法收敛。之后，Wei 和 Tanner（1990）研究了带外生变量的 p 阶删失自回归（删失 ARX(p)），并针对删失数据的条件后验分布提出了一种抽样方案，成功应用了结合数据增广的吉布斯采样器。该程序同样基于 Zeger 和 Brookmeyer（1986）对似然的分解。

本文提出一种贝叶斯方法，用于估计带有 AR(p) 误差的删失回归模型。众所周知，这些模型的系数具有通常的解释方式，因此与 ARX 模型相

比更容易阐释。本文提出的算法采用带数据增广的吉布斯采样器（GDA），在每次迭代中同时对模型参数和潜变量进行抽样。数据增广通过对相应条件分布中的潜变量进行多次抽样来实现。因此，删失观测由多个样本的均值取代，从而使算法收敛更快、估计值更准确。在数据增广的框架下，似然函数的计算可简化为多元高斯样本的似然计算。在时间序列分析中，通常采用条件似然。然而，在当前情形下，通过适当的变量变换可以得到完整似然。此外，本文还提出了一种基于数据增广的贝叶斯框架下的模型选择与模型评估方法。可根据各模型归一化常数之比（即证据）计算贝叶斯因子，以评估竞争模型的相对表现。Friel 和 Wyse（2012）回顾了若干常用的模型证据估计方法。本文进一步为相关研究作出贡献，指出 GDA 可通过预测性能度量（传统上称为信息准则）用于模型选择，并可通过适用于时间序列数据的留一交叉验证进行预测评估。基于合成数据集和真实数据集的实证实验表明，即使删失率较高（40

最后需要指出的是，在方法的构建过程中，本文仅关注左删失。不过，正如将该方法应用于云幕高度时间序列时所示，该方法很容易调整并扩展至右删失情形，这也充分展现了该方法的灵活性。

本文结构安排如下：第2节定义所研究的模型；第3节介绍采用数据增广的贝叶斯方法，详细说明所有必要步骤，并利用合成数据集展示该方法在三种不同删失情景下的表现。第4节讨论使用删失数据时的模型评估；第5节分析一组云幕高度时间序列，该数据此前由 Park 等人（2007）和 Schumacher 等（2017）分析，最初由美国国家大气研究中心（NCAR）收集。该数据集包含1989年3月旧金山的716个每小时观测值，其中41.7%为删失值。结论部分给出一些总结性意见及可能的未来拓展方向。

2 具有自相关误差的删失线性回归

如果潜变量 w 只有高于 L 的值才会被记录，而小于或等于该限值的值均记为 L，则称其在 L 处左删失。观测变量 y 定义为

$$y = \begin{cases} w, \text{ if } w > L \\ L, \text{ if } w \leq L \end{cases} \tag{1}$$

或者等价地， $y = max(w, L)$。类似地，如果 w 发生右删失，记录的值将为 $y = min(w, L)$。 L 可以视为检测限。

现在考虑经典的线性回归模型，其序列相关误差定义为一个 AR(p) 过程，记为 LR-AR。该模型关于时间 t 时响应变量 w_t 的离散时间表示为

$$\begin{aligned} w_t &= \mathbf{x}_t\boldsymbol{\beta} + u_t \\ u_t &= \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_p u_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\epsilon^2) \end{aligned} \tag{2}$$

其中， $\mathbf{x}_t = (1, x_{t2}, \ldots, x_{tk})$ 是一个 $1 \times k$ 维解释变量或特征向量， $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})$ 是 k 维回归系数向量， u_t 是一个具有高斯创新项 ε_t 和 AR 系数 $\boldsymbol{\rho} = (\rho_1 \ldots, \rho_p)$ 的平稳 AR(p) 过程，并满足通常的平稳性条件。

现在假设我们观测到可能被删失的值 $y_t = max(w_t, L)$，其中L 是已知的删失限值。于是，我们将带AR误差的删失线性回归模型 CLR-AR 写为

$$\begin{aligned} y_t &= \max(w_t, L) \\ w_t &= \mathbf{x}_t\boldsymbol{\beta} + u_t, \\ u_t &= \rho_1 u_{t-1} + \rho_2 u_{t-2} + ... + \rho_p u_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \end{aligned} \tag{3}$$

下文中， $\mathbf{y} = (y_1, \ldots, y_T)$ 表示实际记录的值，这些值可能被删失；而 $\mathbf{w} = (w_1, \ldots, w_T)$ 表示相应的潜在过程 w_t， $\mathbf{X}$ 表示回归变量的 $T \times k$ 矩阵， $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\rho}, \sigma_\epsilon^2)$ 表示参数向量。过程 ω_t 在时间 t 之前的历史表示为 $\boldsymbol{\omega}_t = (\omega_1, \ldots, \omega_t)$

3 利用数据增广进行贝叶斯推断

在贝叶斯框架下，对删失数据进行推断的一种自然方法是采用结合数据增广的吉布斯采样器（Gelfand and Smith，1990；Casella and George，1992），即 GDA（Tanner and Wong，1978；Fridley and Dixon，2007；Chib，1992）。该方法可描述为每次迭代中执行的两步程序：(i) 用从截断条件分布中生成的值填补（可能存在的）删失观测，从而得到一个被视为完整的增广数据集；(ii) 从模型参数的全条件分布中生成模型参数。

首先，我们介绍数据增广过程的方法。

3.1 数据增广

通常，数据增广步骤在每次迭代中只对删失观测模拟一个值，这种做法没有考虑截断分布的方差（Hopke 等，2001年）。为克服这一问题，本文提出了一种新的GDA算法：用从截断分布中模拟得到的若干个（例如 m个）值的均值来填补删失观测。对合成数据进行的数值研究（见第3.5节）表明，这种方法称为GDA-MSM，能够得到位置和离散程度良好的参数后验分布。

另一个重要问题涉及用于插补（可能的）删失观测的截断分布。给定一个可能包含删失观测的数据集 $\mathbf{y} = (y_1, ..., y_T)$，增广数据集定义如下：

$$z_t = \begin{cases} y_t, \text{ if } y_t > L \\ z_t \sim F(z_t|\mathbf{y}, \boldsymbol{\theta}, z_t \leq L), \text{ if } y_t = L, \end{cases} \tag{4}$$

其中， $F(z_t|\mathbf{y}\ \boldsymbol{\theta}, z_t \leq L)$ 是潜变量删失值对应的截断分布，其支持集为 $]-\infty, L]$. 具体而言，在高斯假设下

$$f(z_t|\mathbf{y}, \boldsymbol{\theta}, z_t \leq L) = \frac{1}{\sigma} \times \frac{\phi\left(\frac{z_t - \mu_{t|\mathbf{z}_{t-1}}}{\sigma}\right)}{\Phi\left(\frac{L - \mu_{t|\mathbf{z}_{t-1}}}{\sigma}\right)} \times I_{(-\infty, L)}(z_t), \tag{5}$$

其中 $\phi(.)$ 和 $\Phi(.)$ 分别表示标准正态分布的概率密度函数和累积分布函数，并且

$$\begin{aligned} \mu_{t|\,\mathbf{z}_{t-1}} =& \rho_1 z_{t-1} + \rho_2 z_{t-2} + ... + \rho_p z_{t-p} \\ &+ (\mathbf{x}_t - \rho_1 \mathbf{x}_{t-1} - \rho_2 \mathbf{x}_{t-2} - ...\rho_p \mathbf{x}_{t-p})\boldsymbol{\beta}. \end{aligned} \tag{6}$$

所得增广数据向量 $\mathbf{z} = (z_1, \ldots, z_T)$ 被视为潜变量的一组假设观测值，这些观测值满足方程 (2) 中表示的模型，并且是后续贝叶斯分析的对象。以下各节将介绍贝叶斯分析所需的要素：似然和全条件分布。

3.2 完全似然

计算完全似然函数

$$L(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}) = f(z_1, \ldots, z_T|\mathbf{y}, \boldsymbol{\theta}) \tag{7}$$

考虑以下变量变换

$$\mathbf{z}^* = \mathbf{Q}\mathbf{z} \tag{8}$$

其中， $T \times T$ 矩阵 Q 满足 Q′Q 与 u 的方差-协方差矩阵的逆矩阵成比例。事实上，Q 是如下形式的矩阵

$$\mathbf{Q} = \begin{bmatrix} q_{11} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ q_{21} & q_{22} & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{p1} & q_{p2} & \cdots & q_{pp} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ -\rho_p & -\rho_{p-1} & \cdots & -\rho_1 & 1 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & -\rho_p & \cdots & -\rho_2 & -\rho_1 & 1 & \cdots & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\rho_{p-1} & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\rho_p & \cdots & -\rho_1 & 1 \end{bmatrix} \tag{9}$$

其中， q_{ij} 是在式（11）所表达的约束下得到的。事实上，这一变换是由模型（2）中 $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_T)$ 与 $\mathrm{u} = (u_1, u_2, ..., u_T)$ 之间的以下关系所诱导的

$$\boldsymbol{\varepsilon} = \mathbf{Q}\mathbf{u}. \tag{10}$$

令 $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\varepsilon}^2 \mathrm{I}$ 和 $\Sigma_{\mathbf{u}}$ 分别表示 $\boldsymbol{\varepsilon}$ 与 u 的方差—协方差矩阵，且 I 为单位矩阵。由于 $\Sigma_{\boldsymbol{\varepsilon}} = \sigma_{\varepsilon}^2 \mathrm{I} = \mathrm{Q}\Sigma_{\mathbf{u}}\mathrm{Q}'$，因此 Q 必须满足

$$\sigma_{\varepsilon}^2 (\mathbf{Q}'\mathbf{Q})^{-1} = \Sigma_{\mathbf{u}} \tag{11}$$

因此

$$z_t^* = \begin{cases} q_{11} z_1, & t = 1 \\ q_{21} z_1 + q_{22} z_2, & t = 2 \\ \vdots & \\ q_{p1} z_1 + \ldots + q_{pp} z_p, & t = p \\ z_t - \rho_1 z_{t-1} - \ldots - \rho_p z_{t-p}, & t = p+1, ..., T \end{cases} \tag{12}$$

定义 $\mathrm{X}^* = \mathrm{QX}$，得到

$$x_{tj}^* = \begin{cases} q_{11} x_{1j}, & t = 1 \\ q_{21} x_{1j} + q_{22} x_{2j}, & t = 2 \\ \vdots & \\ q_{p1} x_{1j} + \ldots + q_{pp} x_{pj}, & t = p \\ x_{tj} - \rho_1 x_{t-1,j} - \ldots - \rho_p x_{t-p,j}, & t = p+1, ..., T \end{cases} \tag{13}$$

对于 $j = 2$、$\ldots k$ 和 $x_{t1}^* = 1$、 $t = 1$、$\ldots$、T.

因此，似然函数 (7) 等价于

$$L(\mathbf{z}^*|\mathbf{y},\boldsymbol{\theta}) = |\mathbf{Q}|\Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{T}{2}} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(z_t^* - \mathbf{x}_t^*\boldsymbol{\beta})^2\Big\} \tag{14}$$

3.3 全条件分布

贝叶斯分析涉及对先验信息的正式考量，并根据后验分布 $\pi(\boldsymbol{\theta}|\mathbf{y})$ 推断模型参数，该后验分布定义为

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}), \tag{15}$$

其中， $\boldsymbol{\theta}$ 是参数向量， $L(\mathbf{y}|\boldsymbol{\theta})$是观测数据的似然函数，而 $\pi(\boldsymbol{\theta})$表示参数的联合先验分布。在缺乏先验信息时，考虑非信息先验分布，并假设 $\boldsymbol{\beta}$、 σ^2 和 $\boldsymbol{\rho}$ 是相互独立的变量，其先验设定如下

$$\pi(\boldsymbol{\beta}) \propto c_1, \quad \pi(\sigma^2) \propto \frac{1}{\sigma^2}, \quad \pi(\boldsymbol{\rho}) \propto c_2 \times \boldsymbol{I}_{\boldsymbol{\rho}\in S_{\boldsymbol{\rho}}}, \tag{16}$$

其中， c_1、 c_2 是常数， $S_{\boldsymbol{\rho}}$ 是过程 u_t 的平稳区域，而$\boldsymbol{I}_{(\cdot)}$ 表示示性函数。

结合式（14）和式（16），含增广数据的后验分布可写为:

$$\pi(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) \propto |\mathbf{Q}|\Big(\frac{1}{\sigma^2}\Big)^{\frac{T}{2}+1} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(z_t^* - \mathbf{x}_t^*\boldsymbol{\beta})^2\Big\} \times \boldsymbol{I}_{\boldsymbol{\rho}\in S_{\boldsymbol{\rho}}}. \tag{17}$$

由式（17）可得，模型参数的全条件分布为

$$\pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{\rho},\mathbf{y},\mathbf{z}) \propto \exp\Big\{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\frac{1}{\sigma^2}(\mathbf{X}^{*\prime}\mathbf{X}^*)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\Big\}, \tag{18}$$

$$\pi(\sigma^2|\boldsymbol{\beta},\boldsymbol{\rho},\mathbf{y},\mathbf{z}) \propto \Big(\frac{1}{\sigma^2}\Big)^{\frac{T}{2}+1} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(z_t^* - \mathbf{x}_t^*\boldsymbol{\beta})^2\Big\}, \tag{19}$$

$$\pi(\boldsymbol{\rho}|\boldsymbol{\beta},\sigma^2,\mathbf{y},\mathbf{z}) \propto |\mathbf{Q}| \exp\Big\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(z_t^* - \mathbf{x}_t^*\boldsymbol{\beta})^2\Big\} \times \boldsymbol{I}_{\boldsymbol{\rho}\in S_{\boldsymbol{\rho}}}, \tag{20}$$

其中，$\hat{}\,\boldsymbol{\beta} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{z}^*$ 是可行广义最小二乘（FGLS）估计量。式（18）和式（19）的函数形式表明

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{\rho},\mathbf{y},\mathbf{z} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}), \tag{21}$$

$$\sigma^2|\beta,\boldsymbol{\rho},\mathbf{y},\mathbf{z} \sim IG\Big(\frac{T}{2}, \frac{1}{2}(\mathbf{z}^* - \mathbf{X}^*\boldsymbol{\beta})'(\mathbf{z}^* - \mathbf{X}^*\boldsymbol{\beta})\Big). \tag{22}$$

然而，为了对 $\boldsymbol{\rho}$ 的取值进行采样，我们需要在吉布斯采样器中使用 Metropolis-Hastings 算法（Gilks等，1995年）。

3.4 GDA-MSM算法

以下算法说明了如何在 CLR-AR(p) 中使用吉布斯采样器和上述数据增广过程 GDA-MSM 进行贝叶斯推断。

给定数据集 $\mathsf{y} = (y_1, ..., y_T$，其中可能包含删失观测，GDA-MSM算法允许按如下方式为 CLR-AR(p)模型的参数构造马尔可夫链:

算法 1：带数据增广的吉布斯采样器（GDA）

1. 使用 y、 $L \in$R 进行初始化，$N \in$ Z 和 $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \sigma^{2(0)}, \boldsymbol{\rho}^{(0)})$
2. 设置 $\mathsf{z}^{(0)} =$y
3. 对于 $i = 1, ..., N$
4. 样本 $\boldsymbol{\beta}^{(i)} \sim \pi(\boldsymbol{\beta}|\sigma^{2(i-1)}, \boldsymbol{\rho}^{(i-1)}, \mathsf{y}, \mathsf{z}^{(i-1)})$
5. 样本 $\sigma^{2(i)} \sim \pi(\sigma^2|\boldsymbol{\beta}^{(i)}, \boldsymbol{\rho}^{(i-1)}, \mathsf{y}, \mathsf{z}^{(i-1)})$
6. 样本 $\boldsymbol{\rho}^{(i)} \sim \pi(\boldsymbol{\rho}|\boldsymbol{\beta}^{(i)}, \sigma^{2(i)}, \mathsf{y}, \mathsf{z}^{(i-1)})$
7. 对于 $t = 1, ..., T$
8. 如果 $y_t \leq L$
9. 对于 $j = 1, ..., m$
10. 样本 $z_{tj} \sim F(z_t|\mathsf{y}\ \boldsymbol{\theta}^{(i)}, z_t \leq L) \times I_{(z_t \leq L)}$
11. $z_t^{(i)} := \dfrac{1}{m} \sum_{j=1}^{m} z_{tj}$
12. Else
13. $z_t^{(i)} := y_t$
14. 返回 $\boldsymbol{\Theta} = [\boldsymbol{\theta}^{(1)} \cdots \boldsymbol{\theta}^{(N)}]'$和 $\mathsf{z}^{(N)}$。

模型参数的马尔可夫链蒙特卡罗估计值 $\boldsymbol{\theta}$ 通常通过计算 GDA 输出 $\boldsymbol{\Theta}$ 的样本均值来获得，除非边际后验密度显示分布高度偏斜；在这种情况下，使用样本中位数更为合适。所得增广数据 $\mathsf{z}^{(N)} = (z_1^{(N)}, \ldots, z_T^{(N)})$可视为潜变量的观测值，以便进行进一步推断（Tanner and Wong，1978；Law 和 Jackson，2017）。

3.5 合成数据集示例

通过使用从 CLR-AR(1)模型模拟的删失时间序列，可以说明上述方法的性能；模拟数据既包含解释变量，也不包含解释变量。滞后1阶相关取若干正值和负值，即$\rho_1 = -0.8, -0.48, -0.15$、0.15、0.48、0.8；删失比例设为三种情形：5%、20%和40%；样本量则设为100、500和1000。模型参数 $\beta_0, \beta_1, \sigma^2$ 的取值依据 Schumacher 等（2017）和 Wang 和 Chan（2018）的论文确定，具体见表1。注意，标记为M1 的模型对应均值为 $\beta_0/(1 - \rho_1)$的 AR(1)模型。模型总数为十八个，因此共有 $18 \times 3 \times 3 = 162$ 种模拟情形，

这些情形允许控制删失程度和序列相关。

表1：模拟研究中CLR-AR(1)模型的参数 模型 参数

	β_0	β_1	σ^2
M1	2	0	2
M2	2	1	2
M3	0. 2	0.4	0.607

该过程使用R实现（R核心团队，2020），具体而言，使用 $MASS$ （Ripley等人，2021）和 $invgamma$ （Kahle和Stamey，2017）软件包从多元正态分布和逆伽马分布中抽样。算法迭代 $N = 4 \times 10^4$ 次，舍弃了 $b = 2 \times 10^4$ 次初始预热迭代，并且仅保留最后若干次迭代中每隔20个的值，以降低链内的自相关。使用 $Coda$ 软件包（Plummer等人，2006；Robert和Casella，2010）提供的常规诊断检验，对MCMC算法的收敛情况进行了充分分析。初始估计值通过FGLS估计获得，模型参数则根据链中剩余的 $M = 1 \times 10^3$个值计算后验均值。数据增广所模拟的值的数量为 $m = 5$ ，与Hopke等人（2001）所用的数量相同。

图1–3通过在三种删失情景下、 $\rho_1 = 0.48$ 时**M2** 的后验密度，展示了该方法的整体结果和性能。这些图体现了GDA–MMS方法的效率和贝叶斯一致性：对于所考虑的所有删失情景，边际后验分布通常集中在包含参数真实值的集合上（红色竖直虚线），且随着样本量增加，变异性逐渐减小。其他 ρ_1 取值和其他模型的后验密度示例见附录A。

为进一步研究该方法的性质，对162种情景中的每一种分别生成100次实现，结果汇总于附录B的表4至表7中。

因此，该方法在估计具有AR误差的删失回归模型方面效果良好。

4 删失数据中的模型评估

本节介绍贝叶斯分析自相关误差删失回归中的模型评估和模型选择准则。

首先定义 $t+1$ 处的刀切法一步超前残差（Harrison和West,

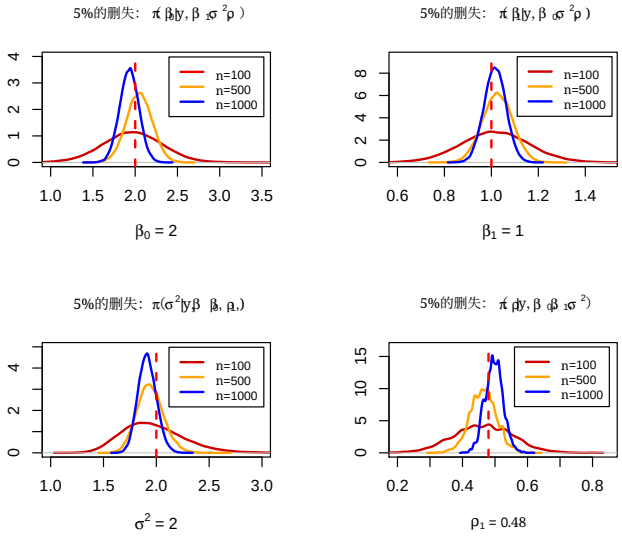

图1：含有 $\rho_1 = 0.48$的模型2：在删失比例为5%的情况下， $n = 100$、500和1000对应的模型参数后验密度。

1991；Shiffrin 等人，2008），即

$$d_{t+1}^{J} = \frac{z_{t+1} - E[z_{t+1}|\mathbf{z}_t]}{\sqrt{Var[z_{t+1}|\mathbf{z}_t]}} \tag{23}$$

这些值通过采用留未来交叉验证（LFO-CV）方法计算得出（Burkner 等人，2020）。该方法是对广受欢迎的留一交叉验证方法的改进：它将所有未来观测值留出，以评估时间序列模型的预测性能。实际操作中，样本 $t = 1, \ldots, T$被划分为训练集和测试集；训练集包含 n 个观测值，并不断扩展，测试集则包含剩余的 $n - T$ 个观测值（Wagenmakers 等人，2006）。选择 n 的值，使得估计具有一致性。

在实际操作中，方程（23）中的均值 $E[z_{t+1}|\mathbf{z}_t]$和方差 $Var[z_{t+1}|\mathbf{z}_t]$ ，通过使用从 z_{t+1}的一步超前预测分布中生成的 M 个值，以其样本对应量近似。

$$f(z_{t+1}|\mathbf{z}_t) = \int_{\Theta} f(z_{t+1}|\mathbf{z}_t, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{z}_t)d\boldsymbol{\theta} \tag{24}$$

其无法用闭式形式表示。因此

$$E[z_{t+1}|\mathbf{z}_t] \approx \frac{1}{M}\sum_{j=1}^{M} z_{t+1}^{(j)}, \tag{25}$$

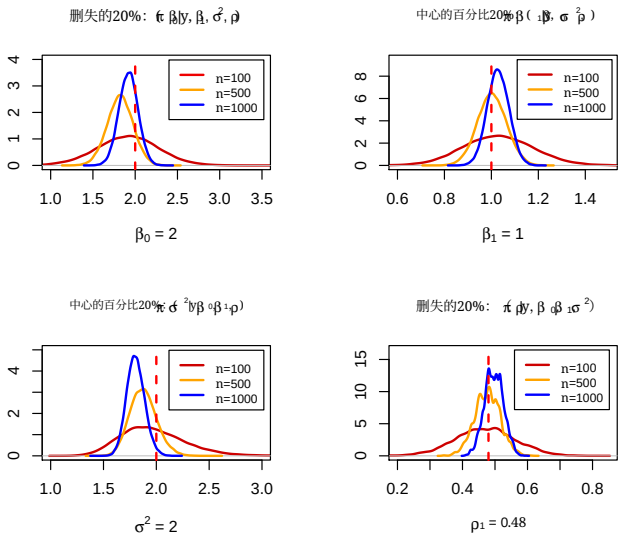

图2：含有 $\rho_1 = 0.48$的模型2：在删失比例为20%的条件下， $n = 100$为500和1000时模型参数的后验密度。

$$Var[z_{t+1}|\mathbf{z}_t] \approx \frac{1}{M-1}\sum_{j=1}^{M}\left(z_{t+1}^{(j)} - E[z_{t+1}|\mathbf{z}_t]\right)^2, \tag{26}$$

其中， $z_{t+1}^{(j)}$从 $N(\mu, \sigma^2)$中模拟得到， μ如方程（6）所示；给定由 $\pi(\boldsymbol{\theta}|\mathbf{z}_t)$生成的马尔可夫链蒙特卡罗输出 $\boldsymbol{\theta}^{(j)}$ 。

在删失数据的情境下，残差 d_t^J（公式（23）、（25）、（26））的计算以及后续的模型评估通过两个步骤完成：

步骤1： 给定（可能经过）删失的数据集 y_1， ....., y_T，通过GDA-MSM算法拟合模型并获得一个增广数据集$\mathbf{z} = (z_1, ..., z_T)$；选择 $n < T$

步骤2： 对于每个 $t = n+1$， ....., T

2.1 将 GDA-MSM算法应用于（$y_1, \ldots, y_t$），生成 $\boldsymbol{\theta}_t^{(j)}$；然后根据方程（6）生成 $z_t^{(j)} \sim N(\mu, \sigma^2)$ μ，其中 $j = 1, \ldots, M$

2.2 使用公式（25）和（26）近似预测分布的期望和方差**2.3** 将 $\mathbf{z} = (z_1, \ldots, z_T)$ 视为实际观测数据，计算（ 23 ）（参见 Gourieroux 等（ 1987 年）、Law 和 Jackson（2017））。

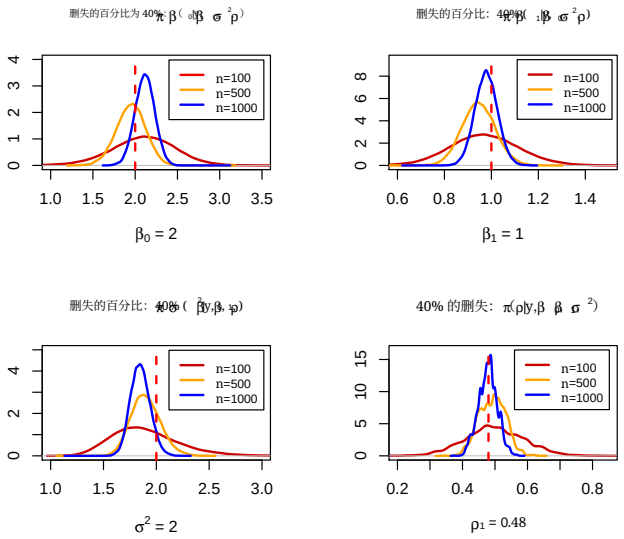

图3：模型2，含 $\rho_1 = 0.48$；在删失比例为40%的条件下， $n = 100$、500和1000的模型参数后验密度。

由此得到的标准化贝叶斯残差（23）现在不仅可用于评估拟合模型的质量，还可通过平方和（或绝对值之和）等指标比较竞争模型的预测性能；在这种情况下，取值较小的模型更受青睐。

关于模型选择，最常用的贝叶斯准则是偏差信息准则DIC（Spiegelhalter等，2002）和广泛适用信息准则WAIC（Watanabe，2013）。DIC和WAIC的表达式见附录C。给定马尔可夫链蒙特卡罗输出，WAIC度量（30）中 pw 的近似值计算如下：

$$pw \approx -2\sum_{t=1}^{T}\left\{\frac{1}{M}\sum_{j=1}^{M} ln f(y_t|\boldsymbol{\theta}^{(j)}) - ln\Big[\frac{1}{M}\sum_{j=1}^{M} f(y_t|\boldsymbol{\theta}^{(j)})\Big]\right\} \tag{27}$$

与残差分析一样，增广数据集z $=(z_1,\ldots,z_T)$用于评估似然函数 $f(\mathrm{y}|\theta)$，方法是将 y_t 替换为通过GDA算法得到的 z_t， $t = 1,\ldots,T$。

5 云底高度时间序列分析

考虑云底高度这一气象时间序列，该序列此前已由 Park 等人（2007）和 Schumacher 等（2017）进行分析。云底高度是

指从地面到云底的距离，以百英尺为单位进行测量。根据 Park 等人（2007）的研究，准确确定云底高度之所以重要，主要是因为它是导致与天气相关的事故的主要因素之一，也是造成航班延误的主要原因之一。记录设备的检测限为 12000 英尺，因此观测数据可视为右删失时间序列。

这些数据最初由美国国家大气研究中心（NCAR）根据1989年3月旧金山的每小时观测数据收集，共有716条观测，其中41.7%为删失观测。对数变换数据可在 R 软件的 $ARCensReg$ 程序包 Schumacher 等人（2016）中获取。图4展示了数据图。

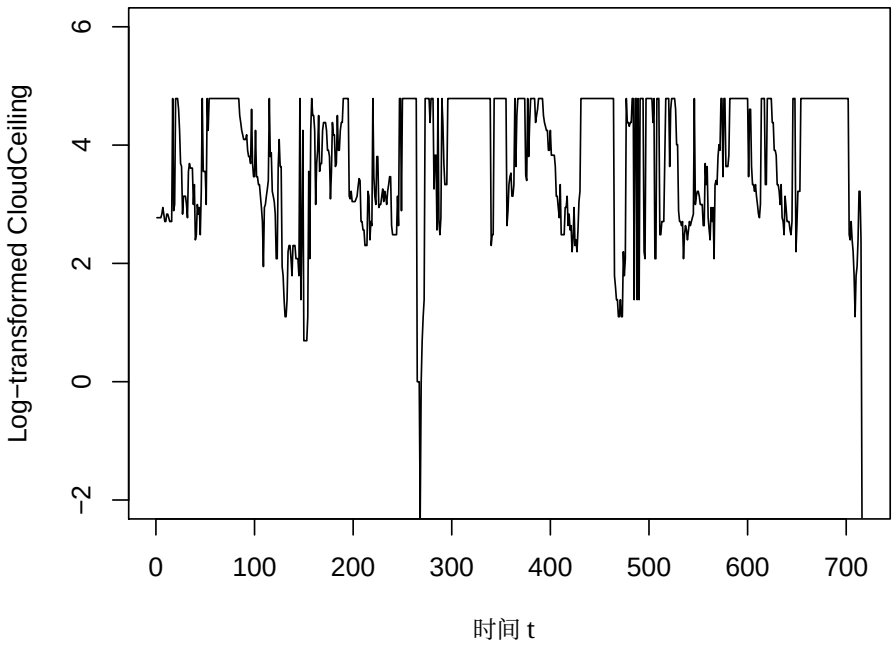

图4：1989年3月旧金山每小时云底高度对数变换数据的删失时间序列

在没有解释变量的情况下，CLR-AR(p) 模型对应于删失 AR(p) 模型，并考虑了 $p = 1,2$ 的两个取值。

根据对链收敛情况的监测（附录 D），CLR-AR(1) 的预热期设为 $b = 2 \times 10^4$，CLR-AR(2) 的预热期设为 $b = 4 \times 10^4$。舍弃预热期后，Geweke 检验（Geweke（1992））的 z 分数为

$p = 1$ 和 $p = 2$分别为 {-0.302, 0.350, 0.624} 和 {0.272, 0.137, 0.787, -0.991}，表明链已收敛（见附录 D）。为减少马尔可夫链蒙特卡罗输出中的自相关，并获得长度为 $M = 1 \times 10^3$ 的子样本以计算估计值，对于 $p = 1$， $lag = 80$ 和 $N = 1 \times 10^5$ 分别设为；对于 $p = 2$，这些值为： $lag = 180$ 和 $N = 2.2 \times 10^5$。附录 D 中的自相关函数（ACF）图表明，这些子样本中没有显著的自相关。参数的后验密度见图5。

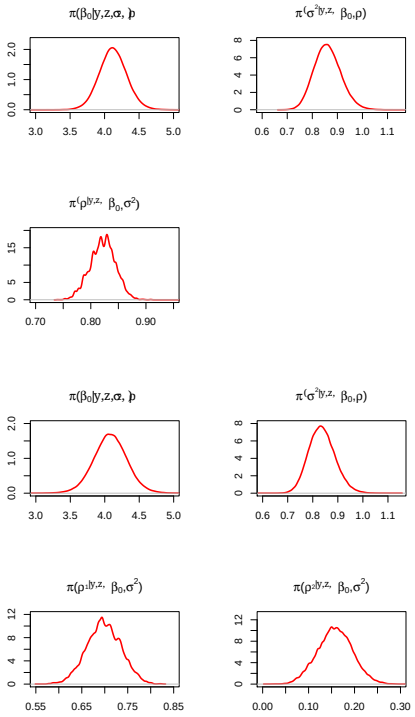

图5：后验密度，上：AR(1)模型；下：AR(2)模型。

为计算参数估计，计算了保留的马尔可夫链蒙特卡罗子样本的样本均值。此外，表2还提供了这些子样本的其他汇总统计量，即中位数、标准误（SE）和置信概率为0.95的HPD可信区间（CI）。CI′是使用 R package H $DInterval$（Meredith和Kruschke，2018）计算的。

为获得刀切法预测残差，初始训练样本的大小设为 $n = 600$ ，相应的均值和方差分别为：AR(1)模型为0.007和1.257，AR(2)模型为0.061和0.95。这些数值分别接近0和1，其图形

图 6上部强调，这些残差分布在零值附近，且没有显著相关性（图 6下部）。

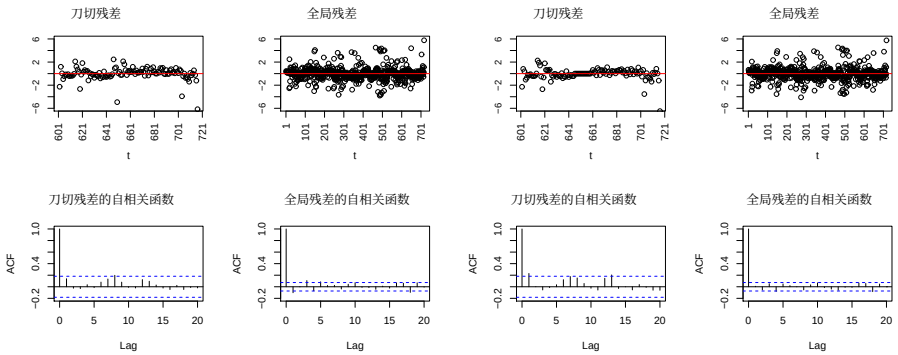

图 6: （左）残差及其对应的自相关函数: （左） $p = 1$（右）$p = 2$.

表 2: 对经对数变换的云底高度数据使用带 AR(p)误差项的 LR模型所得的参数估计。

p	统计量	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\rho}_1$	$\hat{\rho}_2$
1	Mean	4.136	0.862	0.822	–
	中位数	4.134	0.858	0.823	–
	SE	0.205	0.053	0.022	–
	CI	[3.771, 4.543]	[0.768, 0.970]	[0.776, 0.862]	–
2	Mean	**4.070**	**0.839**	**0.697**	**0.160**
	中位数	4.068	0.836	0.695	0.160
	SE	0.248	0.053	0.037	0.037
	CI	[3.620, 4.576]	[0.738, 0.942]	[0.626, 0.766]	[0.087, 0.231]

偏差信息准则、广泛适用信息准则和标准化刀切残差平方和（SSJR）的值列于表3。由于采用 AR(2)误差的模型在偏差信息准则、广泛适用信息准则和 SSJR 方面的值最低，因此选择 AR(2)模型。该结论以及模型参数的取值与 Schumacher 等（2017）分析该数据集时得到的结果一致。

与 AR(2) 误差模型估计值相对应的增广数据见图7（蓝线），并与观测数据（红线）进行对比。

表3：模型评估的信息准则

模型	DIC	WAIC	SSJR
CLR-AR(1)	684.1	416497.4	144.6
CLR-AR(2)	590.6	377259	109.8

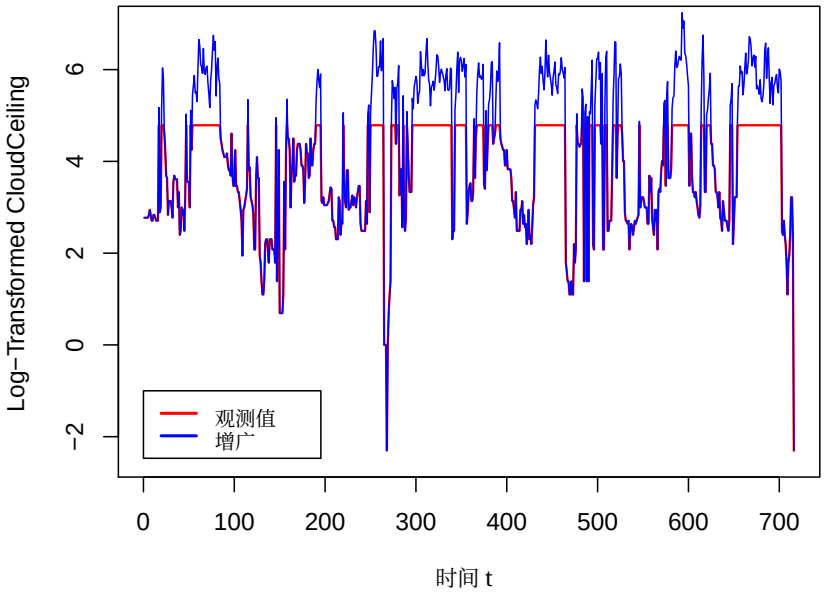

图7：1989年3月旧金山每小时云底高度对数变换后删失时间序列的观测数据与增广数据。

6 结论

本文提出一种贝叶斯方法，用于对含 AR(p) 误差的删失数据线性回归模型（CLR-AR(p) 模型）进行推断。忽略数据中的删失模式并采用常规估计方法会导致估计值有偏。本文提出的算法采用结合数据增广的吉布斯采样器。其创新之处在于采用多次模拟的均值进行数据增广（GDA-MMS），从而提高算法的准确性。事实上，即使删失值的比例很高（40%），GDA-MMS 算法仍能良好运行。

请注意，在模拟和实证示例中使用了杰弗里先验。不过，如果有关于数据的信息，也可以采用具有适当超参数的其他先验；特别是，对于 (β, σ^2)，可以考虑多元正态-逆伽马分布。

这里假设删失阈值已知。未来研究中有待解决的一个问题是，在删失水平未知的情况下对数据进行建模。

致谢

本研究得到卡洛斯特·古尔本基安基金会和数学与应用研究发展中心（CIDMA）的支持，并通过葡萄牙科学技术基金会（FCT - 葡萄牙科学与技术基金会）获得资助，项目编号 UIDB/04106/2020。

参考文献

Beach, C., MacKinnon, J., 1978. 二阶自回归误差模型的完全最大似然估计。计量经济学期刊 7, 187–198. Buckley, J., James, I., 1979. 删失数据的线性回归。Biometrika 66, No. 3, 429–436. Burkner,P-C., Gabry, J., Vehtari, A. 2020 贝叶斯时间序列模型的近似未来留出交叉验证。统计计算与模拟杂志 90, No. 14, 2499–2523. Casella, G., George, E., 1992. 解释吉布斯采样器。美国统计学家 46, No. 3, 167–174. Chib, S., 1992. Tobit删失回归模型中的贝叶斯推断。计量经济学期刊 51, 79–99. Fridley, B., Dixon, P., 2007. 涉及删失观测的贝叶斯空间模型的数据增广。Environmetrics 18, 107—123. Friel, N., Wyse, J., 2012. 估计统计证据——综述。Sta- tistica Neerlandica 66, 288–308. Gelfand, A.E., Smith, A.F.M., 1990. 计算边际密度的基于抽样的方法。美国统计学会杂志 85, No. 410, 398–409. Geweke, J., 1992. 评估基于抽样的方法计算后验矩的准确性（含讨论），载于：Bernardo, J., Berger, J., Dawid, A., , Smith, A. (编)，贝叶斯统计学4。牛津大学出版社，牛津，第 169–193 页。

Gilks, W.R., Best, N.G., Tan, K.K.C., 1995. 吉布斯抽样中的自适应拒绝 Metropolis 抽样。皇家统计学会期刊 44, 4, 455–472. Gourieroux, C., Monfort, A., Renault, E., Trognon, A., 1987. 模拟残差。计量经济学期刊 34, 201–252. Harrison, J., West, M., 1991. 动态线性模型诊断。Bio metrika 78, 4, 797–808. Hopke, P., Liu, C., Rubin, D., 2001. 对含缺失值和低于阈值测量值的多变量数据进行多重插补：北极污染物的时间序列浓度。Biometrics 57, 22–33. Houseman, E.A., Virji, M.A., 2017. 对左删失的时间序列暴露数据进行汇总和建模的贝叶斯方法。Annals of Work Ex- posures and Health 61, No. 7, 773—783. Kahle, D., Stamey, J., 2017. R 软件包’ invgamma’：逆伽马分布。CRAN 仓库。Law, M., Jackson, D., 2017. 含删失结果数据的线性回归模型残差图：一种用于直观呈现残差不确定性的改进方法。统计学通讯——模拟与计算 46:4, 3159–3171. Meredith, M., Kruschke, J., 2018. hdinterval软件包。CRAN 仓库，1–7. Mohammad, N.M., 2014. 删失时间序列分析。博士论文。西安大略大学，安大略省。Park, J., Genton, M., Ghosh, S., 2007. 使用自回归移动平均模型进行删失时间序列分析。加拿大统计学杂志 35, 1, 151–168. Plummer, M., Best, N., Cowles, K., Vines, K., 2006. Coda：马尔可夫链蒙特卡罗的收敛诊断与输出分析。R News 6, 7–11. Prais, S., Winsten, C., 1954. 趋势估计量与序列相关。Cowles 委员会讨论论文：统计学，第 383 号。R核心团队，2020. R：统计计算语言与环境。R统计计算基金会。奥地利维也纳。Ripley, B., Venables, B., Hornik, K., Gebhardt, A., Firth, D., 2021. R 软件包’mass’：Venables 和 Ripley 的 MASS 的辅助函数与数据集。CRAN 仓库。Robert, C., Casella, G., 2010. R语言蒙特卡罗方法导论。施普林格，纽约。

Schumacher, F., Lachos, V., Galarza, C., 2016. R 包 'arcensreg'：拟合具有自回归误差的单变量删失线性回归模型。CRAN 仓库。Schumacher, F.L., Lachos, V., Dey, D., 2017. 具有自回归误差的删失模型：基于似然的视角。《加拿大统计学杂志》45, 68, 375–392。Shiffrin, R.M., Lee, M.D., Kim, W., Wagenmakers, E.J., 2008. 模型评估方法综述及层次贝叶斯方法教程。《认知科学》32, 1248–1284。Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. 模型复杂度与拟合度的贝叶斯度量。《英国皇家统计学会会刊》14, 867–897。Tanner, M., Wong, W., 1978. 通过数据增广计算后验分布。《美国统计学会会刊》82, 第 398 期, 528–540。Wagenmakers, E.J., Gruwald, P., Steyvers, M., 2006. 累积预测误差与时间序列模型的选择。《数学心理学杂志》50, 149–166。Wang, C., Chan, K., 2018. 含外生变量的删失自回归模型的准似然估计。《美国统计学会杂志》113:523, 1135–1145。Watanabe, S., 2013. 一种广泛适用的贝叶斯信息准则。《机器学习研究杂志》14, 867–897。Wei, G.C.G., Tanner, M.A., 1990. 删失回归数据的后验计算。《美国统计学会会刊》85, 829–839。Zangari, P., Tsurumi, H., 1996. 对日本乘用车出口数据中删失自相关数据的贝叶斯分析。《计量经济学进展》11, A 部分, 111–143。Zeger, S., Brookmeyer, R., 1986. 含删失自相关数据的回归分析。《美国统计学会杂志》81, 722–729。

A 后验密度

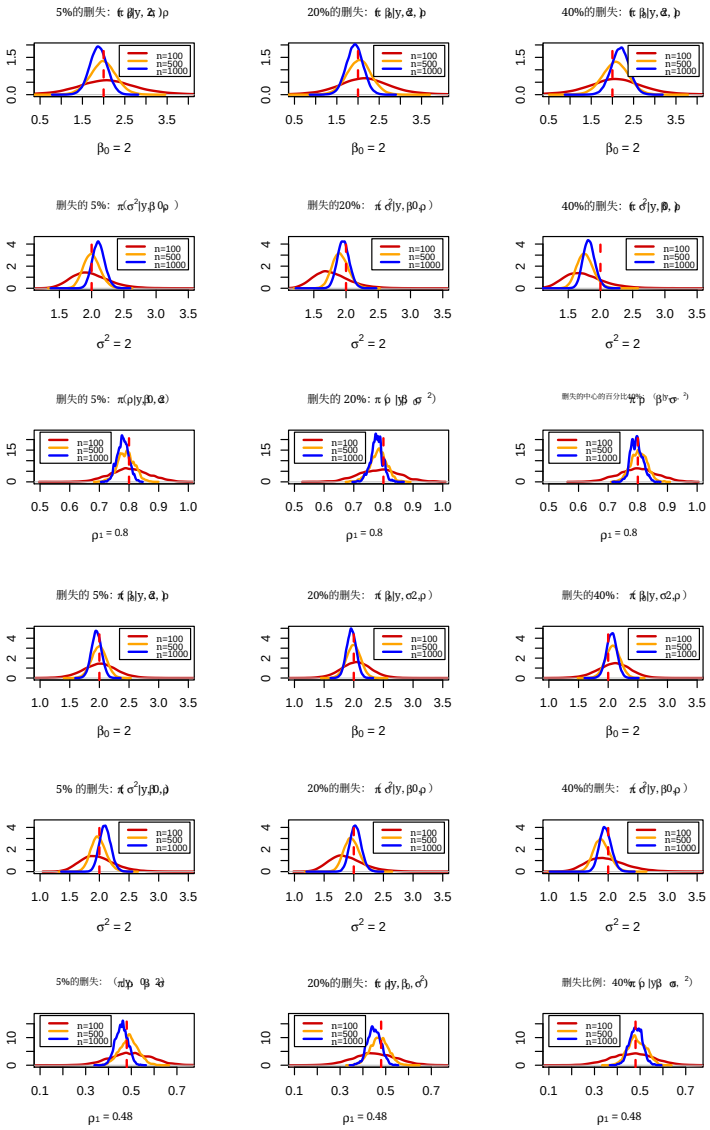

图8：模型 *M1* 的 $\rho = 0.8$ 上方 3 条线和 $\rho = 0.48$ 下方 3 条线。在 3 种删失情景下，$n = 100$、500 和 1000 的模型参数后验密度。

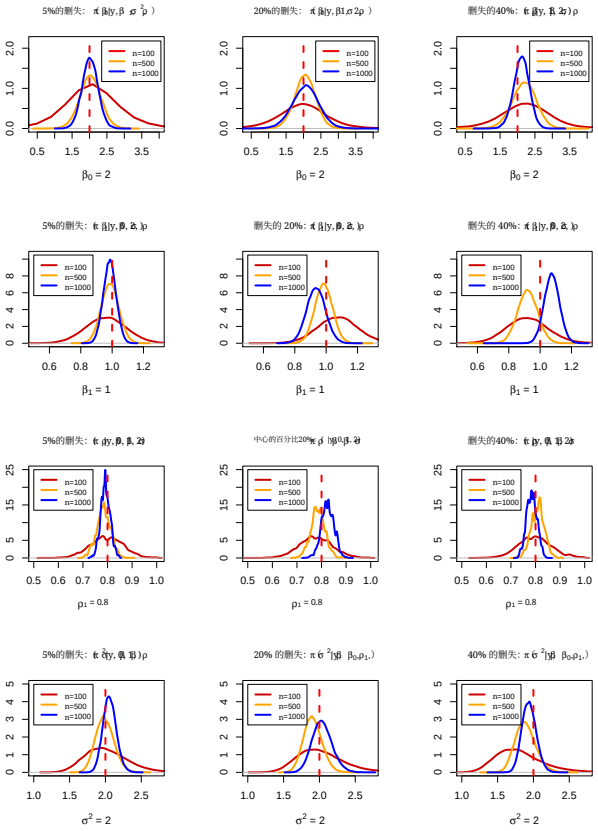

图9：包含 $\rho = 0.8$的模型M2：在3种删失情景下， $n = 100$、500和1000对应的模型参数后验密度。

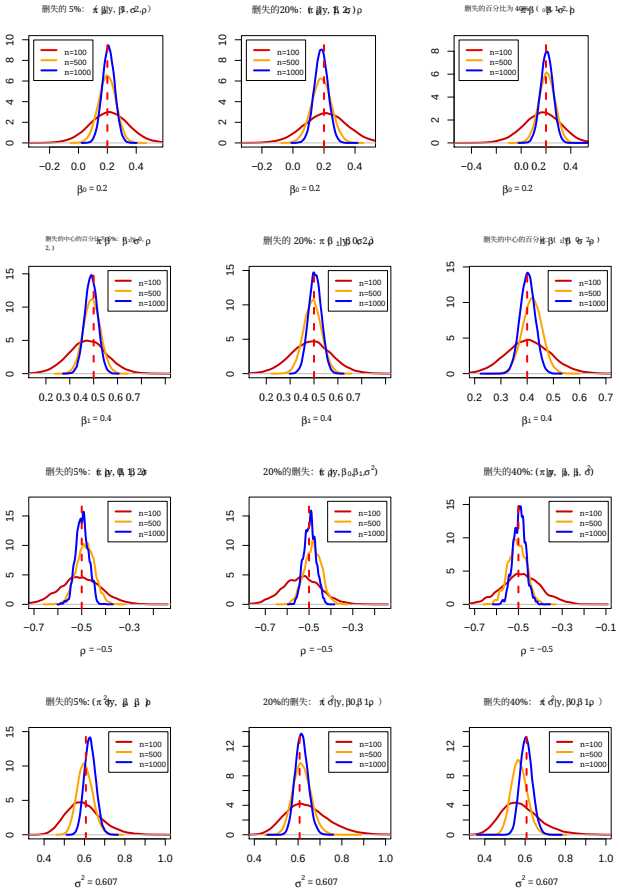

图10：模型 *M3*，包含 $\rho = -0.5$：在 3 种删失情景下，$n = 100$ 为 500 和 1000 时模型参数的后验密度。

B 模拟结果

表4：**模型2**：结果，均值（标准差），基于在不同样本量、删失率和 $\rho > 0$ 条件下对简单 CLR-AR(1)模型进行的 100 次模拟。

n	中心的百分比	$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\rho = 0.15$
100	5%	2.027(0.066)	0.995(0.026)	2.036(0.107)	0.134(0.011)
	20%	2.025(0.070)	0.995(0.025)	1.962(0.107)	0.133(0.012)
	40%	2.069(0.100	0.980(0.030)	1.756(0.165)	0.127(0.016)
500	5%	1.996(0.018)	1.006(0.006)	2.014(0.017)	0.148(0.002)
	20%	2.009(0.018)	1.002(0.006)	1.921(0.025)	0.147(0.002)
	40%	2.109(0.026)	0.968(0.007)	1.659(0.135)	0.140(0.003)
1000	5%	2.007(0.009)	0.995(0.003)	2.004(0.009)	0.149(0.001)
	20%	2.026(0.011)	0.987(0.003)	1.911(0.017)	0.145(0.001)
	40%	2.119(0.025)	0.956(0.005)	1.659(0.125)	0.141(0.001)
n	中心的百分比	$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\rho = 0.48$
100	5%	2.046(0.085)	0.992(0.023)	2.045(0.106)	0.460(0.008)
	20%	2.045(0.098)	0.992(0.024)	1.985(0.120)	0.452(0.010)
	40%	2.098(0.119)	0.983(0.028)	1.767(0.179)	0.450(0.012)
500	5%	2.001(0.024)	1.006(0.006)	2.015(0.017)	0.477(0.002)
	20%	2.020(0.027)	1.000(0.007)	1.929(0.023)	0.472(0.002)
	40%	2.125(0.038)	0.966(0.008)	1.696(0.115)	0.463(0.002)
1000	5%	2.009(0.013)	0.995(0.003)	2.004(0.009)	0.477(0.001)
	20%	2.031(0.016)	0.987(0.003)	1.924(0.016)	0.471(0.001)
	40%	2.130(0.032)	0.957(0.006)	1.689(0.108)	0.462(0.001)
n	中心的百分比	$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\rho = 0.8$
100	5%	2.110(0.354)	0.992(0.019)	2.106 (0.105)	0.780(0.004)
	20%	2.130(0.369)	0.999(0.019)	1.991(0.111)	0.777(0.004)
	40%	2.290(0.397)	0.968(0.027)	1.755(0.214)	0.762(0.007)
500	5%	2.022(0.099)	1.005(0.005)	2.016(0.017)	0.793(0.001)
	20%	2.054(0.101)	1.000(0.005)	1.942(0.022)	0.789(0.001)
	40%	2.200(0.119)	0.979(0.007)	1.714(0.107)	0.779(0.002)
1000	5%	2.013(0.058)	0.995(0.002)	2.005(0.009)	0.795(0.000)
	20%	2.054(0.101)	1.000(0.005)	1.942(0.022)	0.789(0.001)
	40%	2.200(0.119)	0.979(0.007)	1.714(0.107)	0.779(0.002)

表5：**模型2**：基于不同样本量和删失率下简单 CLR-AR(1)模型的100次模拟结果，使用 GDA 和多个样本的均值以及 $\boldsymbol{\rho < 0}$。

n	中心的百分比	$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\rho = -0.15$
100	5%	2.013(0.056)	1.000(0.023)	2.041(0.109)	-0.161(0.009)
	20%	2.006(0.060)	1.003(0.023)	1.960(0.098)	-0.160(0.012)
	40%	2.066(0.076	0.980(0.024)	1.760(0.162)	-0.160(0.014)
500	5%	1.994(0.015)	1.007(0.006)	2.014(0.017)	-0.152(0.002)
	20%	2.011(0.016)	1.000(0.006)	1.914(0.026)	-0.151(0.002)
	40%	2.090(0.023)	0.974(0.006)	1.672(0.125)	-0.151(0.003)
1000	5%	2.005(0.007)	0.997(0.002)	2.005(0.009)	-0.151(0.001)
	20%	2.011(0.016)	1.000(0.006)	1.917(0.025)	-0.151(0.002)
	40%	2.119(0.025)	0.956(0.005)	1.655(0.128)	-0.148(0.001)
n	中心的百分比	$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\rho = -0.48$
100	5%	2.003(0.045)	1.004(0.017)	2.055(0.115)	-0.481(0.007)
	20%	1.992(0.057)	1.009(0.018)	1.989(0.102)	-0.479(0.008)
	40%	2.054(0.065)	0.985(0.018)	1.837(0.123)	-0.467(0.009)
500	5%	1.992(0.011)	1.007(0.004)	2.021(0.017)	-0.479(0.002)
	20%	2.003(0.011)	1.003(0.004)	1.919(0.030)	-0.475(0.002)
	40%	2.090(0.023)	0.972(0.005)	1.741(0.086)	-0.474(0.002)
1000	5%	2.001(0.005)	0.999(0.002)	2.011(0.009)	-0.479(0.001)
	20%	2.012(0.006)	0.995(0.002)	1.940(0.014)	-0.476(0.001)
	40%	2.097(0.018)	0.965(0.004)	1.740(0.078)	-0.468(0.001)
n	中心的百分比	$\beta_0 = 2$	$\beta_1 = 1$	$\sigma^2 = 2$	$\rho = -0.8$
100	5%	1.992(0.033)	1.007(0.012)	2.089 (0.120)	-0.797(0.003)
	20%	1.994(0.052)	1.003(0.015)	2.097(0.134)	-0.793(0.003)
	40%	2.065(0.103)	0.969(0.029)	2.059(0.141)	-0.780(0.005)
500	5%	1.991(0.010)	1.007(0.003)	2.033(0.018)	-0.793(0.001)
	20%	1.998(0.010)	1.007(0.003)	2.025(0.021)	-0.794(0.001)
	40%	2.071(0.023)	0.977(0.006)	1.947(0.030)	-0.786(0.001)
1000	5%	1.999(0.004)	1.000(0.001)	2.019(0.009)	-0.797(0.000)
	20%	1.992(0.010)	1.006(0.003)	2.025(0.021)	-0.794(0.001)
	40%	2.061(0.013)	0.977(0.003)	1.959(0.016)	-0.784(0.001)

表6：**模型1**：在不同样本量和删失比例下，使用 GDA、多个样本的均值和 $\boldsymbol{\rho > 0}$，基于无解释变量的 CLR-AR(1)模型进行100次模拟所得的结果。

n	中心的百分比	$\beta_0 = 2$	$\sigma^2 = 2$	$\rho = 0.15$
100	5%	2.021(0.021)	2.037(0.108)	0.133(0.010)
	20%	2.021(0.021)	1.974(0.111)	0.131(0.011)
	40%	2.066(0.023)	1.693(0.208)	0.125(0.014)
500	5%	2.006(0.058)	2.015(0.017)	0.149(0.002)
	20%	2.011(0.006)	1.931(0.023)	0.147(0.002)
	40%	2.053(0.008)	1.681(0.120)	0.142(0.003)
1000	5%	2.000(0.003)	2.005(0.009)	0.148(0.001)
	20%	2.005(0.003)	1.923(0.016)	0.148(0.001)
	40%	2.046(0.004)	1.673(0.116)	0.143(0.001)
n	中心的百分比	$\beta_0 = 2$	$\sigma^2 = 2$	$\rho = 0.48$
100	5%	2.034(0.055)	2.044(0.105)	0.461(0.008)
	20%	2.037(0.055)	1.993(0.135)	0.456(0.008)
	40%	2.087(0.056)	1.753(0.195)	0.444(0.012)
500	5%	2.010(0.015)	2.015(0.016)	0.477(0.002)
	20%	2.018(0.015)	1.946(0.023)	0.473(0.002)
	40%	2.067(0.017)	1.725(0.096)	0.462(0.002)
1000	5%	2.001(0.008)	2.001(0.009)	0.477(0.001)
	20%	2.008(0.008)	1.938(0.014)	0.472(0.001)
	40%	2.058(0.010)	1.717(0.090)	0.461(0.001)
n	中心的百分比	$\beta_0 = 2$	$\sigma^2 = 2$	$\rho = 0.8$
100	5%	2.097(0.368)	2.056(0.111)	0.781(0.004)
	20%	2.126(0.346)	2.000(0.122)	0.774(0.004)
	40%	2.266(0.341)	1.708(0.231)	0.762(0.006)
500	5%	2.027(0.097)	2.018(0.019)	0.793(0.001)
	20%	2.054(0.095)	1.939(0.021)	0.788(0.001)
	40%	2.191(0.109)	1.673(0.133)	0.778(0.002)
1000	5%	2.003(0.054)	2.009(0.009)	0.795(0.000)
	20%	2.036(0.053)	1.926(0.016)	0.789(0.001)
	40%	2.157(0.065)	1.682(0.115)	0.780(0.001)

表7：**模型1**：基于不同样本量和删失率下对不含解释变量的 CLR-AR(1)模型进行的100次模拟结果，使用GDA和多个样本的均值以及 $\boldsymbol{\rho < 0}$。

n	中心的百分比	$\beta_0 = 2$	$\sigma^2 = 2$	$\rho = -0.15$
100	5%	2.017(0.012)	2.035(0.107)	-0.163(0.010)
	20%	2.011(0.013)	1.972(0.104)	-0.162(0.011)
	40%	2.051(0.015)	1.721(0.174)	-0.169(0.013)
500	5%	2.004(0.003)	2.017(0.018)	-0.152(0.002)
	20%	2.008(0.003)	1.932(0.023)	-0.152(0.002)
	40%	2.046(0.005)	1.688(0.115)	-0.154(0.003)
1000	5%	2.000(0.002)	2.004(0.009)	-0.151(0.001)
	20%	2.004(0.002)	1.925(0.015)	-0.150(0.001)
	40%	2.045(0.004)	1.674(0.115)	-0.152(0.001)
n	中心的百分比	$\beta_0 = 2$	$\sigma^2 = 2$	$\rho = -0.48$
100	5%	2.013(0.007)	2.048(0.112)	-0.487(0.007)
	20%	2.008(0.009)	1.996(0.114)	-0.486(0.008)
	40%	2.026(0.013)	1.850(0.139)	-0.478(0.008)
500	5%	2.003(0.002)	2.021(0.018)	-0.480(0.002)
	20%	2.007(0.003)	1.954(0.021)	-0.478(0.002)
	40%	2.047(0.005)	1.714(0.100)	-0.481(0.002)
1000	5%	2.000(0.001)	2.008(0.009)	-0.479(0.001)
	20%	2.004(0.001)	1.947(0.013)	-0.476(0.001)
	40%	2.040(0.003)	1.755(0.070)	-0.471(0.001)
n	中心的百分比	$\beta_0 = 2$	$\sigma^2 = 2$	$\rho = -0.8$
100	5%	2.002(0.005)	2.068(0.108)	-0.790(0.003)
	20%	1.993(0.007)	2.051(0.118)	-0.787(0.004)
	40%	2.012(0.013)	1.953(0.122)	-0.783(0.004)
500	5%	2.000(0.001)	2.029(0.019)	-0.797(0.001)
	20%	2.003(0.002)	1.984(0.020)	0.796(0.001)
	40%	2.029(0.004)	1.882(0.039)	-0.793(0.001)
1000	5%	1.999(0.001)	2.015(0.010)	-0.797(0.000)
	20%	2.001(0.001)	1.982(0.011)	-0.795(0.000)
	40%	2.025(0.002)	1.885(0.025)	-0.791(0.000)

C 预测性能的度量

DIC

偏差信息准则是贝叶斯方法中广泛使用的一种拟合优度度量，计算如下:

$$DIC = -4 \cdot E_{\theta|\mathbf{y}}[lnf(\mathbf{y}|\theta)] + 2 \cdot lnf(\mathbf{y}|\hat{\theta}), \tag{28}$$

其中， $E_{\theta|\mathbf{y}}[lnf(\mathbf{y}|\theta)]$ 是对数似然函数的后验均值，定义为

$$E_{\theta|\mathbf{y}}[lnf(\mathbf{y}|\theta)] = \frac{1}{M}\sum_{j=1}^{M} lnf(\mathbf{y}|\boldsymbol{\theta}^{(j)}), \tag{29}$$

ˆ而 $f(\mathbf{y}|\theta)$ 是在贝叶斯参数估计值处计算的似然函数。

WAIC

广泛适用信息准则是另一种预测准确度度量，与先前的准则相比，它与贝叶斯方法的联系更为密切（Watanabe，2013），其定义为

$$WAIC = -2\sum_{t=T}^{n} lnE_{\theta|\mathbf{y}}[f(y_t|\theta)] + 2pw. \tag{30}$$

其中， pw 是校正项，通常采用偏差信息准则（DIC）中使用的校正项，定义如下:

$$pw = -2\sum_{t=1}^{T}\{E_{\theta|\mathbf{y}}[lnf(y_t|\theta)] - lnE_{\theta|\mathbf{y}}[f(y_t|\theta)]\}. \tag{31}$$

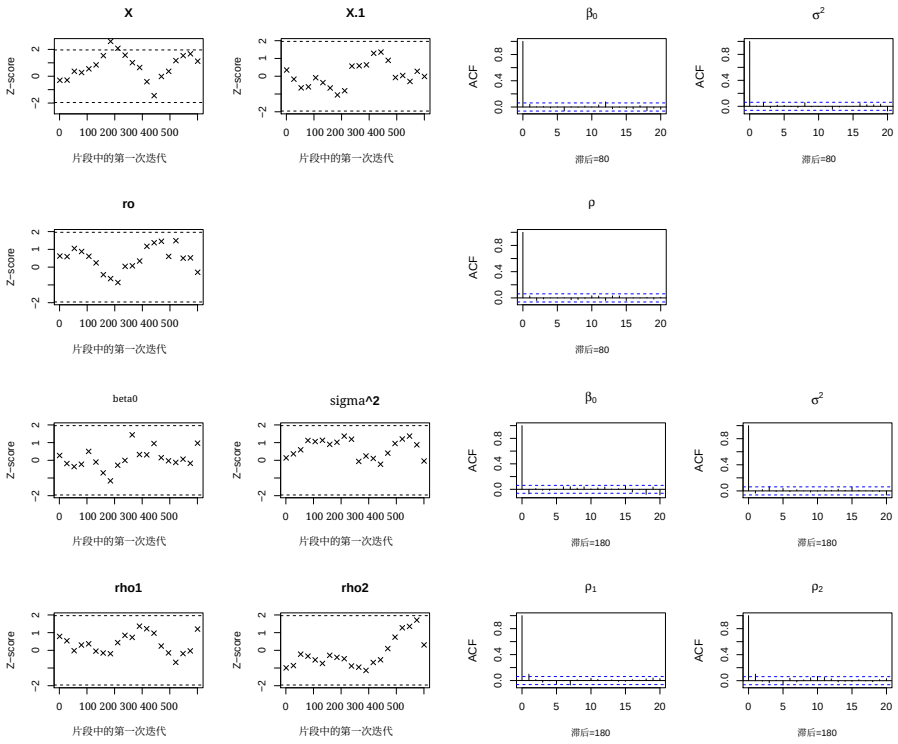

图11：上：用于计算 CLR-AR(1) 参数估计值的子样本的 Geweke图和自相关函数。下：用于计算 CLR-AR(2) 参数估计值的子样本的 Geweke图和自相关函数。

D 链的收敛性分析

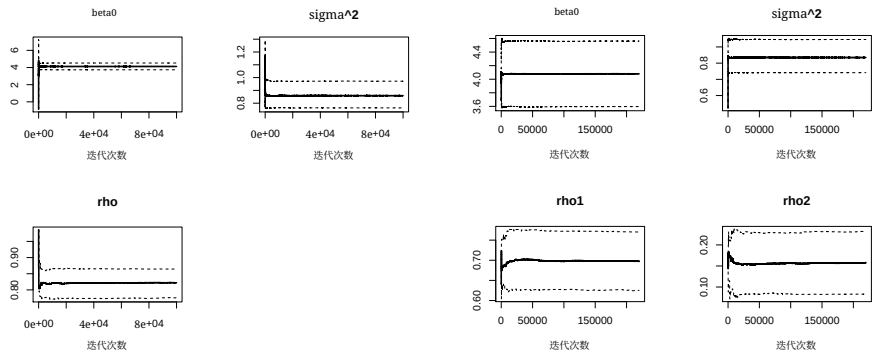

图12：马尔可夫链蒙特卡罗输出对应的第1和第3分位数的变化，上方：AR(1)模型；下方：AR(2)模型。