arXiv:2603.03987v1 [stat.ME] 4 Mar 2026

针对膨胀有界数据的贝叶斯结构化加性分位数回归

Francisco F. Queiroz [1], JohannesBrachem [2], Paul F.V. Wiemann [2], ThomasKneib [2],

[1]圣保罗大学统计学系，巴西[2]哥廷根大学统计学教席，德国[2]俄亥俄州立大学统计学系，美国

摘要

单位区间上的有界连续数据常见于应用领域，并且边界处往往有不可忽略比例的观测值。膨胀回归模型通过将单位区间上的连续分布与离散分量相结合，处理零值和／或一值膨胀现象。本文提出一类适用于膨胀有界连续数据的贝叶斯结构化加性分位数回归模型，可处理零值和／或一值膨胀。该方法能够直接对连续部分的条件分位数以及观测到零值和／或一值的概率进行建模，并在这两个部分中纳入结构化加性预测变量，包括非线性效应、空间效应、随机效应和变系数项。后验推断采用马尔可夫链蒙特卡罗算法完成，并通过用于半参数回归的概率编程框架软件 Liesel 实现。我们通过模拟研究和两项真实数据应用展示所提模型的实际表现：一项分析巴西市辖区交通相关死亡人数所占比例，另一项评估不同实验条件下人工耳蜗植入者的言语可懂度。

关键词：有界连续数据、分位数回归、分布回归、半参数回归 *n.*

1 引言

单位区间上的有界连续数据常见于医学、生物学、社会学和生态学等领域，例如死亡率、体脂百分比以及树冠覆盖的土地比例。这类数据往往在边界处包含不可忽略数量的观测值，即大量的零值和／或一值。例如，某种疾病的死亡率或感染率、用于子女教育的家庭收入比例，以及具有某种特征的人口比例。对这类数据进行建模的一种自然方法是采用膨胀回归模型，其中响应变量 y 服从混合分布，该分布由一个取值于 (0，1) 上的连续随机变量和一个用于表示边界值的离散随机变量构成。零膨胀和一膨胀随机变量的概率密度函数（PDF）通常可表示为

$$g(y; p_0, p_1, \boldsymbol{\vartheta}) = \left\{ p_0^{\mathbb{I}_{\{0\}}(y)} p_1^{\mathbb{I}_{\{1\}}(y)} (1 - p_0 - p_1)^{\mathbb{I}_{(0,1)}(y)} \right\} \left\{ h(y; \boldsymbol{\vartheta})^{\mathbb{I}_{(0,1)}(y)} \right\}, \quad y \in [0, 1], \tag{1}$$

其中 $\mathbb{I}_A(y)$为示性函数，当 $y \in A$ 时等于1，否则等于0； $h(y; \boldsymbol{\vartheta})$表示可能依赖于参数向量 $\boldsymbol{\vartheta}$的连续随机变量的概率密度函数， $0 < p_0, p_1 < 1$ 和 $p_0 + p_1 < 1$。

*通讯作者：Francisco F. Queiroz，电子邮箱：felipeq@ime.usp.br。

参数 p_0 和 p_1 分别表示零值概率和取值为一的概率。零膨胀或一膨胀随机变量的概率密度函数可由公式（1）通过设定 $p_1 = 0$得到零膨胀情形，或设定 $p_0 = 0$ 得到一膨胀情形。注意，概率密度函数 g 可分解为两项：第一项仅依赖于 p_0 和 p_1 （混合分布离散部分的参数），第二项仅依赖于 $\boldsymbol{\vartheta}$ （混合分布连续部分的参数）。因此，我们称参数（p_0， p_1）与 $\boldsymbol{\vartheta}$ 是可分离的，基于似然的推断可对离散部分和连续部分独立进行。此外，注意公式（1）中的第一项对应于单次试验且概率向量为（p_0， p_1， $1 - p_0 - p_1$）的多项随机变量的概率函数，此时所考虑的变量为（$\mathbb{I}_{\{0\}}(y)$ $\mathbb{I}_{\{1\}}(y)$ $\mathbb{I}_{(0,1)}(y)$）。若 p_0 或 p_1 等于零，则该项退化为伯努利分布的概率函数。

文献中已有多种模型被提出，用于对膨胀有界连续数据进行建模，其中响应变量的形式如方程（1）所示。具体而言，若 $h(y;\boldsymbol{\vartheta})$ 是定义在 $(0,1)$上的某一连续分布的概率密度函数，则可据此构建相应的有界数据零膨胀和/或一膨胀回归模型。此类模型包括基于贝塔分布的模型（Ospina & Ferrari，2010年，2012年；Stasinopoulos 等人，2024年）、基于单位韦布尔分布的模型（Menezes 等人，2021年），以及基于幂逻辑分布的模型（Queiroz & Ferrari，2024年）等。在这些模型中，由于离散成分和连续成分的参数向量彼此可分，因而似然函数也可分，通常会分别对两者的参数进行基于似然的推断和贝叶斯推断。我们将这种独立处理两个成分的建模策略称为两部分建模框架。Kosmidis 和 Zeileis（2025）最近提出了另一种处理边界观测值的方法。作者引入了扩展支撑贝塔（XBX）回归模型，在标准贝塔回归的基础上进行扩展，使其在边界处也能具有正概率。若有理由认为边界观测值和连续观测值来自同一个潜在数据生成过程，而非来自不同的离散和连续机制，那么这种方法尤其具有吸引力。

许多回归模型通过有限的一组参数来刻画响应变量的条件分布，通常关注均值和精度等一两个特定方面。虽然这类以参数为核心的模型可能非常成功，但它们只允许协变量效应作用于条件分布的有限方面，并引入了参数分布假设。分位数回归提供了一种更灵活的视角，它直接针对条件分位数，从而能够进行更全面的分析，捕捉均值趋势之外的异质性、非对称性和协变量效应，且无需作出分布假设。分位数回归由 Koenker 和 Bassett 提出（1978年），并已应用于各种领域；例如，参见 Koenker（2004）、Kneib（2013）、Razen 等（2023）及其中的参考文献。已有一些研究探讨了如何使用分位数回归方法对区间 $(0,1)$ 内的有界数据建模。例如，Bottai 等（2010年）基于逻辑变换提出了用于有界连续数据的分位数回归，而 Bayes 等（2017）则提出了一种采用 Kumaraswamy 分布的参数化分位数混合回归方法。最近，Bourguignon 等（2024年）提出了一种基于广义三参数贝塔分布的参数分位数回归模型。

针对有界数据的膨胀回归模型，涉及分位数回归的研究较少。一些例子见于Santos和Bolfarine（2015年）以及Menezes 等人（2021年），其中提出了分位数回归模型，用于处理零膨胀或一膨胀有界连续数据。Menezes 等人（2021年）所采用的方法与Bourguignon等人（2024年）的方法相平行，利用单位韦布尔分布中基于分位数的参数化，为模型的连续部分建立分位数回归。Santos和Bolfarine（2015年）提出的方法与Koenker和Bassett（1978年）最初引入的分位数回归概念十分相近。作者提出贝叶斯分位数回归模型来解释连续部分的条件分布，同时将混合概率建模为协变量的函数。他们将非对称拉普拉斯分布用作辅助误差分布。

Santos和Bolfarine（2015年）和Menezes 等人（2021年）提出的模型对预测变量采用简单的线性回归设定；当数据在两个边界处均包含观测值时，也就是存在零膨胀和一膨胀时，这些模型并不适用。本文针对膨胀有界连续数据，提出了更灵活的分位数回归模型，并在贝叶斯框架下估计其参数。

Thes这些模型为捕捉膨胀有界数据的分布特征提供了一种灵活的方法 a.

本研究的主要贡献有两点。首先，所提出的模型纳入了结构加性回归，能够包含各种结构化效应，例如协变量的光滑函数、空间效应、随机效应和变系数项。其次，这些模型旨在处理零值或一值膨胀以及零值和一值膨胀的有界数据。据我们所知，目前尚无现有的分位数回归模型能够处理零值和一值膨胀数据。此外，本研究提出的方法为扩展到其他类别的膨胀模型提供了灵活的基础。

本文其余部分安排如下。第 2 节介绍用于零值和/或一值膨胀有界连续数据的贝叶斯结构化加性分位数回归模型，同时讨论两部分建模框架、结构化加性预测变量，以及基于非对称拉普拉斯分布的贝叶斯推断。第 3节介绍后验抽样方案。第4节展示模拟研究。最后，第 5 节通过两个涉及交通相关死亡和语音可懂度实验的真实数据应用，说明所提出模型的实际用途。文章最后讨论主要发现及未来研究方向。

2 膨胀有界数据的贝叶斯结构化加性分位数回归模型

我们提出一种针对零膨胀和/或一膨胀数据的贝叶斯半参数结构化加性分位数回归模型，并将其纳入两部分建模框架。第一部分通过对边界点（零、一或两者）处的膨胀概率进行建模，处理离散分量；第二部分则侧重于连续部分，刻画有界响应变量的条件分位数。在不失一般性的前提下，我们假设响应变量定义在单位区间上，因为所提出的方法可以很容易地推广到边界已知的任意有界区间。

为形式化零膨胀和一膨胀结构化加性分位数回归模型，令 $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$ 表示第 i 个个体的响应变量 y_i 及其对应协变量向量 $\boldsymbol{x}_i = (x_{1i}, ..., x_{qi})^\top$ 的观测值。我们假设给定协变量后，各响应值条件独立。响应变量 y_i 取值于集合 $[0, 1]$, , 允许在两个边界点零和一处发生膨胀。令 $p_{0i} = \mathbb{P}(y_i = 0)$ 和 $p_{1i} = \mathbb{P}(y_i = 1)$ 分别表示响应值为零和一的概率，其中 $p_{0i} \in (0, 1)$、 $p_{1i} \in (0, 1)$，且对所有 $i = 1, \ldots, n$，均有 $p_{0i} + p_{1i} < 1$。令 $Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)$ 表示在区间 $(0, 1)$ 内观测值中，给定 $\boldsymbol{x}_i$ 时 y_i 的第 $i = 1, \ldots, n$ 条件分位数。由于响应变量被限制在一个有界区间内，我们不直接对 $Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)$ 建模，而是对 $Q_{h(y_i)|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)$ 建模，其中 $h : (0, 1) \to \mathbb{R}$ 是严格递增的链接函数。我们依据以下事实

$$Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i) = h^{-1}\left[Q_{h(y_i)|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)\right]$$

以便在原始尺度上获得响应变量的条件分位数。这种方法使我们能够使用标准分位数回归技术，对实数轴上的条件分位数进行建模。我们不假设膨胀概率 p_i 和条件分位数 $Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)$的线性预测子仅为简单形式，而是采用具有结构化加性预测变量的灵活半参数回归框架（Fahrmeir 等，2021年；Kneib 等，2023；Waldmann 等，2013年），其一般形式为

$$d_0\left(\frac{p_{0i}}{p_{2i}}\right) = \eta_{0i} = \beta_{00} + \sum_{k=1}^{m_0} f_{0k}(\boldsymbol{x}_i), \tag{2}$$

$$d_1\left(\frac{p_{1i}}{p_{2i}}\right) = \eta_{1i} = \beta_{10} + \sum_{k=1}^{m_1} f_{1k}(\boldsymbol{x}_i), \tag{3}$$

$$h\left[Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)\right] = \eta_{2i,\tau} = \beta_{20,\tau} + \sum_{k=1}^{m_2} f_{2k,\tau}(\boldsymbol{x}_i), \tag{4}$$

其中，$d_0, d_1 : (0,1) \to \mathbb{R}$ 是严格单调且二阶可微的连接函数，β_{00}, β_{10} 和 $\beta_{20,\tau}$ 是截距，$f_{0k}(\boldsymbol{x}_i)$、$f_{1k}(\boldsymbol{x}_i)$ 和 $f_{2k,\tau}(\boldsymbol{x}_i)$ 表示用于捕捉不同类型回归效应的一般函数，每个函数均取决于完整协变量向量 $\boldsymbol{x}_i$的某些子集。对于 d_0 和 d_1，自然的选择是与多项logit参数化对应的对数连接函数；对于 h，自然的选择是logit连接，定义为 $\text{logit}(x) = \log[x/(1-x)]$, $x \in (0,1)$。为简化符号，在以下描述中省略β_{20}、τ、η_{2i}、τ中对 τ的依赖，以及函数$f_{2k,\tau}(\boldsymbol{x}_i)$中的相关依赖，并分别简记为 β_{20}、η_{2i}和 $f_{2k}(\boldsymbol{x}_i)$。

常见效应的一些示例如下：

i. 线性：$f(\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$，其中 $\boldsymbol{\beta}$ 是参数向量，它取决于第 τ 个分位数；

ii. 光滑半参数分量：$f(\boldsymbol{x}_i) = f_{\text{smooth}}(x_{ri})$，其中 x_{ri} 是 $\boldsymbol{x}_i$ 的连续协变量，f_{smooth} 是满足某些条件的光滑函数；

iii. 空间分量：$f(\boldsymbol{x}_i) = f_{\text{spatial}}(\boldsymbol{s}_i)$，其中 $\boldsymbol{s}_i$ 是空间位置变量向量；

iv. 变系数项：$f(\boldsymbol{x}_i) = x_{ri} f_{\text{smooth}}(x_{si})$，其中 x_{ri} 是连续或二元协变量，其效应受 x_{si}（连续协变量）调节。

方程（2）—（4）中的结构化加性预测变量的每个非线性效应均通过基函数展开表示为

$$f_{jk}(\boldsymbol{x}_i) = \sum_{l=1}^{L_{jk}} \beta_{jkl} B_{jkl}(\boldsymbol{x}_i),$$

对于 $j = 0, 1, 2$，其中 $B_{jkl}(\boldsymbol{x}_i)$，对于 $l = 1, \ldots\ldots L_{jk}$，是依赖于特定效应类型的基函数，$\beta_{jkl}$ 是其对应系数。用矩阵记号，预测变量可写为

$$\boldsymbol{\eta}_j = \beta_{j0}\mathbf{1}_n + \boldsymbol{B}_{j1}\boldsymbol{\beta}_{j1} + \cdots + \boldsymbol{B}_{jm_j}\boldsymbol{\beta}_{jm_j}, \tag{5}$$

对于 $j = 0, 1, 2$，其中 $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \ldots\ldots, \beta_{jkL_{jk}})^\top$ 是基系数向量，$\boldsymbol{B}_{jk}$ 是由在协变量处求值的基函数得到的设计矩阵，即 $\boldsymbol{B}_{jk}[i,l] = B_{jkl}(\boldsymbol{x}_i)$。详见 Fahrmeir 等（2021年，第8–9章）。

在两部分建模框架中，离散分量通过多项回归模型加以设定，其中因变量，即 $(\mathbb{I}_{\{0\}}(y_i)\ \mathbb{I}_{\{1\}}(y_i)\ \mathbb{I}_{(0,1)}(y_i))$，$i = 1, \ldots\ldots, n$，假定在给定协变量的条件下条件独立，并服从单次试验、概率向量为 (p_{0i}, p_{1i}, p_{2i}) 的多项分布。连续部分对集合 $\mathcal{I} = \{i : y_i \in (0,1)\}$ 中观测值的变换响应变量 $h(y_i)$ 的条件分位数进行建模。在不对 $h(y_i)$ 假定完整分布的情况下，分位数回归的估计可通过最小化非对称加权绝对偏差（非对称加权绝对偏差）之和来实现，通常借助线性规划技术；详见 Koenker（2005）。然而，对于贝叶斯推断，非对称拉普拉斯分布（非对称拉普拉斯分布）因其便利的位置-尺度混合表示而被广泛用作工作模型，这大大简化了后验分布计算（Kozumi 与 Kobayashi，2011年；Santos 与 Bolfarine，2015年；Waldmann 等，2013）。因此，我们假定 $y_i^\dagger := h(y_i)$ 服从非对称拉普拉斯分布，其位置参数 η_{2i} 如式（4）所设定，精度参数 δ^2，偏度参数 τ，概率密度函数由下式给出

$$g_{y_i^\dagger}(y; \eta_{2i}, \delta^2, \tau) = \tau(1-\tau)\delta^2 \exp\{-\delta^2 \rho_\tau(y - \eta_{2i})\}, \quad y \in \mathbb{R},$$

其中，$\rho_\tau(v) = \tau|v|$ 若 $v \geq 0$ 且 $\rho_\tau(v) = (1-\tau)|v|$ 否则。在这一假设下，τ的第 $h(y_i)$ 分位数为 η_{2i}。采用非对称拉普拉斯分布得到的后验众数与非对称加权绝对偏差之和的最小化解一致；参见 Waldmann 等（2013年）。

我们考虑 Yue 和 Rue（2011）提出的非对称拉普拉斯分布的位置-尺度混合表示。

命题 2.1（非对称拉普拉斯分布的位置-尺度混合表示）。设 $z \sim N(0,1)$ 和 $w \sim Exp(\delta^2)$ 是两个相互独立的随机变量，分别服从标准正态分布和率参数为 δ^2 的指数分布。那么，

$$y^{\dagger} = \eta + \xi w + \sigma z\sqrt{\frac{w}{\delta^2}} \sim ALD(\eta, \delta^2, \tau),$$

其中 $\xi = (1-2\tau)/[\tau(1-\tau)]$ 以及 $\sigma^2 = 2/[\tau(1-\tau)]$。

因此，一旦对 w 进行插补，贝叶斯推断模型就成为带有偏移量 ξw 和权重$\sigma\sqrt{w/\delta^2}$的条件高斯回归。

基于两部分建模框架，我们对观测模型作出如下设定：

$$\left(\mathbb{I}_{\{0\}}(y_i), \mathbb{I}_{\{1\}}(y_i), \mathbb{I}_{(0,1)}(y_i)\right) | \beta_{00}, \boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{1m_0}, \beta_{10}, \boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{1m_1} \sim \text{Multinomial}(1; p_{0i}, p_{1i}, p_{2i}),$$

对于 $i = 1, \ldots, n$，以及

$$h(y_i)|\beta_{20}, \boldsymbol{\beta}_{21}, \ldots, \boldsymbol{\beta}_{2m_2}, w_i, \delta^2 \sim \text{N}(\eta_{2i} + \xi w_i, w_i\sigma^2/\delta^2), \quad i \in \mathcal{I},$$

其中 $w_i|\delta^2 \sim \text{Exp}(\delta^2)$。我们假设其先验服从伽马分布 Gamma$(a_0, b_0)$，其中 $a_0 = b_0 = 0.001$，并将其作为δ^2的先验。为了对方程（2）-（4）中的估计函数施加特定性质，为每个系数向量$\boldsymbol{\beta}_{jk}$ 指定一个密度如下的秩亏多元正态先验分布：

$$\pi(\boldsymbol{\beta}_{jk}|\nu_{jk}^2) \propto \left(\frac{1}{\nu_{jk}^2}\right)^{\text{rank}(\mathbf{K}_{jk})/2} \exp\left\{-\frac{1}{2\nu_{jk}^2}\boldsymbol{\beta}_{jk}^{\top}\mathbf{K}_{jk}\boldsymbol{\beta}_{jk}\right\}, \tag{6}$$

其中 $j = 0$、1、2、 $k = 1$、...、 m_j、$\mathbf{K}_{jk}$是半正定且可能秩亏的惩罚矩阵，其选择旨在使效应具有所需的正则化性质； ν_{jk}^2 是先验平滑方差，用于确定先验对后验分布的影响。假设 ν_{jk}^2服从逆伽马超先验，即$\nu_{jk}^2 \sim \text{IG}(a_{jk}, b_{jk})$，并采用弱信息先验设定 $a_{jk} = b_{jk} = 0.01$。 $\boldsymbol{\beta}_{jk}$ 的先验设定可直接与惩罚似然框架联系起来，在该框架中，每个效应都要接受形式为 $\text{pen}(\boldsymbol{\beta}_{jk}) = \lambda_{jk}\boldsymbol{\beta}_{jk}^{\top}\mathbf{K}_{jk}\boldsymbol{\beta}_{jk}$的二次惩罚。其中， λ_{jk} 是一个平滑参数，其作用类似于 ν_{jk}^{-2} 。关于如何为不同类型的效应构建设计矩阵和惩罚矩阵，详见 Fahrmeir 等 (2021年，第9章)；另请参见 Waldmann等人 (2013年) 和 Klein 等 (2015年)。

假设所有先验分布彼此独立，则未归一化后验分布为

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto &\left\{\prod_{i=1}^{n}\left[p_{0i}^{\mathbb{I}_{\{0\}}(y_i)}p_{1i}^{\mathbb{I}_{\{1\}}(y_i)}p_{2i}^{\mathbb{I}_{(0,1)}(y_i)}\right]\right\}\left[\prod_{k=1}^{m_0}\pi(\boldsymbol{\beta}_{0k}|\nu_{0k}^2)\pi(\nu_{0k}^2)\right]\left[\prod_{k=1}^{m_1}\pi(\boldsymbol{\beta}_{1k}|\nu_{1k}^2)\pi(\nu_{1k}^2)\right] \\ &\times\left[\prod_{i\in\mathcal{I}}\phi(y_i^{\dagger}; \eta_{2i} + \xi w_i, w_i\sigma^2/\delta^2)\pi(w_i|\delta^2)\right]\left[\prod_{k=1}^{m_2}\pi(\boldsymbol{\beta}_{2k}|\nu_{2k}^2)\pi(\nu_{2k}^2)\right]\pi(\delta^2),\end{aligned}$$

其中， $\boldsymbol{\theta}$汇集了所有模型参数， $\boldsymbol{y} = (y_1, \ldots, y_n)^{\top}$，而 $\phi(;\, a, b)$是均值为 a 、方差为 b的正态分布的概率密度函数。

单侧膨胀。零值或一值膨胀结构加性分位数回归模型是零膨胀和一膨胀结构化加性分位数回归模型的极限情形，在 $p_{0i} \to 0$ （产生一膨胀）或 $p_{1i} \to 0$ （产生零膨胀）时得到。在此设定下，响应变量 y_i 取值于集合$\{c\} \cup (0, 1)$，其中 $c = 0$ 或 $c = 1$，具体取决于情形；离散分量通过二元回归模型指定，其中因变量，即 $\mathbb{I}_{\{c\}}(y_i)$、 $i = 1$、...、 n，假定在给定协变量的条件下相互独立，并服从成功概率为 $p_i = \mathbb{P}(y_i = c)$的伯努利分布。对于零膨胀，令 $p_i = p_{0i}$，模型预测变量由方程（2）–（4）给出，其中 $p_{2i} = 1 - p_{0i}$。类似地，对于一膨胀，令 $p_i = p_{1i}$，模型预测变量由方程（3）–（4）给出，其中 $p_{2i} = 1 - p_{1i}$。因此，当离散部分采用对数链接函数时，所得离散模型对应于结构化加性逻辑斯蒂回归模型，其预测变量为$\text{logit}(p_i)$。其余

模型设定（包括先验分布和推断程序）均可由一般的零膨胀和一膨胀形式类推得到。为完整起见，零值或一值膨胀结构加性分位数回归模型中 $\boldsymbol{\theta}$ 的未归一化后验分布为

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left\{\prod_{i=1}^{n}\left[p_i^{\mathbb{I}_{\{c\}}(y_i)}(1-p_i)^{1-\mathbb{I}_{\{c\}}(y_i)}\right]\right\}\left[\prod_{k=1}^{m_c}\pi(\boldsymbol{\beta}_{ck}|\nu_{ck}^2)\pi(\nu_{ck}^2)\right]$$
$$\times\left[\prod_{i\in\mathcal{I}}\phi(y_i^{\dagger};\eta_{2i}+\xi w_i, w_i\sigma^2/\delta^2)\pi(w_i|\delta^2)\right]\left[\prod_{k=1}^{m_2}\pi(\boldsymbol{\beta}_{2k}|\nu_{2k}^2)\pi(\nu_{2k}^2)\right]\pi(\delta^2),$$

其中 $\boldsymbol{\theta}$ 汇集了所有模型参数。

注释 2.1。 如果预测变量中的所有函数都是线性的，我们便得到 *Santos* 和 *Bolfarine* 提出的零膨胀或一膨胀分位数回归模型（*2015*年）。

我们以两点关于所提出的建模框架的说明结束本节。首先，尽管所有模型参数的联合后验分布均已指定，但与离散分量相关的参数不依赖于分位数水平 τ，因此可以只估计一次，并用于所有分位数水平。相比之下，连续部分的参数因分位数而异，因此必须针对每个 τ 单独估计。其次，由于连续部分仅针对 (0，1) 内的观测值定义，因此其估计效应应解释为给定 $y_i \in$ (0，1) 时响应变量的条件分位数。

3 后验抽样

我们采用马尔可夫链蒙特卡罗 (MCMC) 方法，从零膨胀和一膨胀结构化加性分位数回归模型参数的后验分布中抽取样本。我们通过迭代地从全条件后验分布中抽样，以分块方式抽取样本。对于回归系数，我们采用 Metropolis-Hastings 步骤，并使用经过调整的迭代加权最小二乘提议分布，这与 Klein 等（2015年）类似；另见 Gamerman（1997年）。提议分布为多元正态分布，

$$\boldsymbol{\beta}_{jk}^{*} \sim N(\mathbf{b}_{jk}, \epsilon_{jk}^2\mathbf{F}(\boldsymbol{\beta}_{jk})^{-1}), \quad \text{with} \quad \mathbf{b}_{jk} = \boldsymbol{\beta}_{jk} + [\epsilon_{jk}^2/2]\mathbf{F}(\boldsymbol{\beta}_{jk})\mathbf{s}(\boldsymbol{\beta}_{jk}), \tag{7}$$

其中 $\mathbf{F}(\boldsymbol{\beta}_{jk})$ 和 $\mathbf{s}(\boldsymbol{\beta}_{jk})$ 分别表示在 $\boldsymbol{\beta}_{jk}$处计算的对数全条件后验分布的负海森矩阵和梯度，而 ϵ_{jk} 是一个步长，在预热阶段通过双重平均算法（Hoffman 和 Gelman，0.8 2014年）调节，使接受概率达到 2014年）。负海森矩阵和梯度通过 JAX（DeepMind 等，2020年）的自动微分计算。

平滑方差 ν_{jk}^2 通过 Gibbs 更新进行采样，从逆伽马全条件分布中抽取

$$\nu_{jk}^2 \mid \cdot \sim \mathrm{IG}\big(a_{jk} + 0.5\mathrm{rank}(\mathbf{K}_{jk}),\ b_{jk} + 0.5\boldsymbol{\beta}_{jk}^{\top}\mathbf{K}_{jk}\boldsymbol{\beta}_{jk}\big).$$

为了获得 w_i的样本，我们使用 ALD 的位置-尺度混合表示。由此可知，给定 δ^2时， w_i独立同分布，且 $w_i|\delta^2 \sim$ Exp(δ^2)，其中 $i \in \mathcal{I}$。因此， w_i^{-1} 的全条件分布为逆高斯分布，即

$$w_i^{-1}|\cdot \sim \text{InvGauss}\left(\sqrt{\frac{\xi^2+2\sigma^2}{(y_i^{\dagger}-\eta_{2i})^2}}, \frac{\delta^2(\xi^2+2\sigma^2)}{\sigma^2}\right), \quad i \in \mathcal{I}.$$

更多细节请参见 Waldmann 等（2013年）。我们从该全条件分布中对 w_i^{-1} 进行 Gibbs 更新。其余模型参数 w_1、……、w_n和 δ则在对数尺度上使用通用无转向采样器（NUTS，Hoffman 和 Gelman，2014年）进行更新。

这些模型在 Liesel 中实现，Liesel 是一个专为半参数回归设计的概率编程框架（Riebl 等人，2023）。协变量效应的基矩阵和惩罚矩阵在内部从

R 中的 mgcv 包获取（Wood，2017；Wood，2025）。所提模型实现的更多细节见补充材料第 1 节。说明所提模型实现的完整代码可在 GitHub 仓库 inflated quantreg 中找到。

4 数值结果

我们开展了一项模拟研究，以评估所提出的膨胀有界数据结构化加性分位数回归方法的实现，并考察使用非对称拉普拉斯分布对连续部分建模时似然错设的影响。这里，我们展示零膨胀和一膨胀情况下的结果；对于仅零膨胀或仅一膨胀的情况，也观察到了类似的结果。

我们根据零膨胀和一膨胀回归模型生成数据，离散分量的设定如下：

$$\log\left(\frac{p_{0i}}{p_{2i}}\right) = 0.3 \times [m_1(x_{1i}) + m_2(x_{2i})] - k,$$
$$\log\left(\frac{p_{1i}}{p_{2i}}\right) = 0.3 \times [m_3(x_{1i}) + m_4(x_{2i})] - k,$$

对于 $i = 1, \ldots, n$，其中 $m_1(x) = \sin(3\pi x)e^{-x}$，$m_2(x) = x^3$，$m_3(x) = 0.5\exp(-x^2) - 0.2$，以及 $m_4(x) = 1$，$p_{2i} = 1 - p_{0i} - p_{2i}$，常数 k 经过选取，使得观测到 $\{0, 1\}$ 的概率具有两个范围：较低（$p_{0i} + p_{1i} \in (0.15, 0.21)$ $k = 2.5$）和中等（$p_{0i} + p_{1i} \in (0.41, 0.58)$ $k = 1$）。我们考虑两种情形来生成连续部分的数据：

S1. 我们通过对从位置参数为 η_{2i}、精度参数为 $\delta^2 = 9$、偏度参数为 $\tau \in \{0.1, 0.5, 0.9\}$ 的非对称拉普拉斯分布中抽取的数据应用 logit 变换，得到连续数据，其中

$$\eta_{2i} = m_1(x_{1i}) + m_2(x_{2i}).$$

注意，η_{2i} 对应于经过 logit 变换的连续响应变量的第 τ 分位数。因此，由于我们使用非对称拉普拉斯分布来估计连续部分的分位数，这属于设定正确的情形。

S2. 我们从 GJS-t 分布生成连续数据(ϱ)（Lemonte & Baz´an，2016），其中位数为 μ_i、精度为 $\sigma_i = x_{2i}$、自由度为 $\varrho = 4$，其中

$$\text{logit}(\mu_i) = m_1(x_{1i}) + m_2(x_{2i}).$$

GJS-t(ϱ) 分布的第 τ 分位数为

$$y_i^\tau = \frac{\mu_i e^{\sigma_i t^\tau}}{1 - \mu_i(1 - e^{\sigma_i t^\tau})}, \tag{8}$$

其中，t^τ 是自由度为 4 的学生 t 分布的第 τ 分位数。

协变量 x_{1i} 和 x_{2i} 独立地从单位区间上的均匀分布中抽取，并在所有重复样本中保持固定。我们考虑样本量 $n = 200$,500。

我们使用第 2节提出的结构化加性零膨胀和一膨胀分位数回归模型，按照以下结构

$$\log\left(\frac{p_{0i}}{p_{2i}}\right) = \beta_{00} + f_{01}(x_{1i}) + f_{02}(x_{2i}),$$
$$\log\left(\frac{p_{1i}}{p_{2i}}\right) = \beta_{10} + f_{11}(x_{1i}) + f_{12}(x_{2i}),$$
$$\text{logit}\left[Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)\right] = \beta_{20} + f_{21}(x_{1i}) + f_{22}(x_{2i}),$$

对于 $i = 1$、.....、 n。我们考虑 $\tau \in \{0.1, 0.5, 0.9\}$，并进行 100 次重复。按照第3节所述，我们使用 Python 中的 Liesel 库进行后验推断。预热期设为 1,500次迭代，抽取5,000 个后验样本，并运行四条并行链。

为评估结果，我们在覆盖观测协变量值范围的 100 点等距网格上，计算预测函数值及其置信区间。对于每个后验样本，我们计算预测函数的均方误差（MSE），其定义为

$$\text{MSE}\left(\hat{f}^{[j]}(x_i)\right) = \frac{1}{r}\sum_{i=1}^{r}\left[f(x_i) - \hat{f}^{[j]}(x_i)\right]^2, \tag{9}$$

其中 f 和 $\hat{f}^{[j]}(x_i)$ 分别表示第 i个网格点在第 j个后验样本中的真实效应和预测效应，索引 $i = 1, \ldots, r$ 用于标识网格点； $r = 100$。图 1 展示了离散分量中每个协变量效应的均方误差根（RMSE）图，该值是对所有后验样本取平均所得。

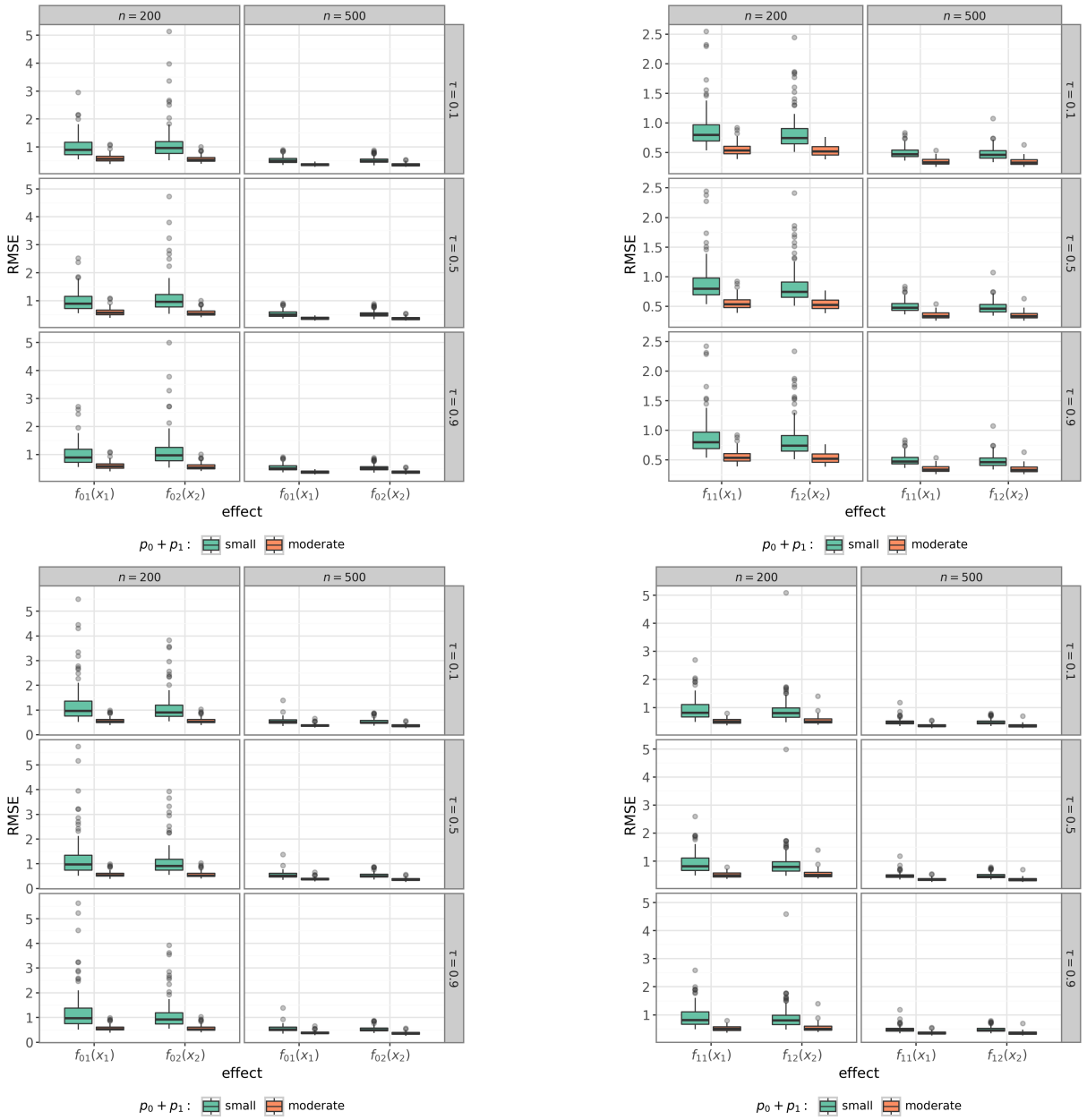


图 1：情景 S1（第一行）和情景 S2（第二行）中离散部分各协变量效应的所有后验样本 RMSE 箱线图。

结果表明，对于较大的样本量（$n = 500$），效应的均方误差根（RMSE）非常接近零。对于中等样本量（$n = 200$），且观测到 $\{0, 1\}$ 的概率较低时，一些 RMSE 会显著增大。这是预期之中的，因为在这种情形下，离散分量的响应变量

$$(\mathbb{I}_{\{0\}}(y_i), \mathbb{I}_{\{1\}}(y_i), \mathbb{I}_{(0,1)}(y_i))$$

主要集中在 (0，0，1)，因此几乎无法提供用于估计离散参数的信息。相比之下，当观测到 $\{0, 1\}$ 的概率适中时，即使对于 $n = 200$，均方根误差仍接近于零。此外，离散分量效应的均方根误差似乎不受所建模分位数选择的影响，这与预期一致。

为评估连续部分的估计效果，我们针对连续型结果的分位数，基于所有后验样本计算均方根误差。y_i 的估计 τ阶条件分位数为

$$\widehat{Q}_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i) = \text{logit}^{-1}\left[\hat{\beta}_{20} + \hat{f}_{21}(x_{1i}) + \hat{f}_{22}(x_{2i})\right].$$

随后，使用新的观测值，根据预测分位数和公式 (8) 中定义的真实分位数计算均方根误差。图2 展示了情景 S1 和 S2 下预测分位数在所有后验样本中的均方根误差。在两种情景下，与极端分位数相比，对中位数分位数建模时的均方根误差更小。当观测到 $\{0, 1\}$ 的概率较小时，均方根误差往往更低，这符合预期，因为此时连续部分的观测值更多。最后，情景 S1 下的均方根误差小于情景 S2，这也符合预期，因为情景 S1 对应于模型设定正确（数据由非对称拉普拉斯分布生成）。尽管如此，情景 S2 下的均方根误差仍然较低且接近于零，表明即使在似然错设的情况下，模型仍能很好地拟合数据。

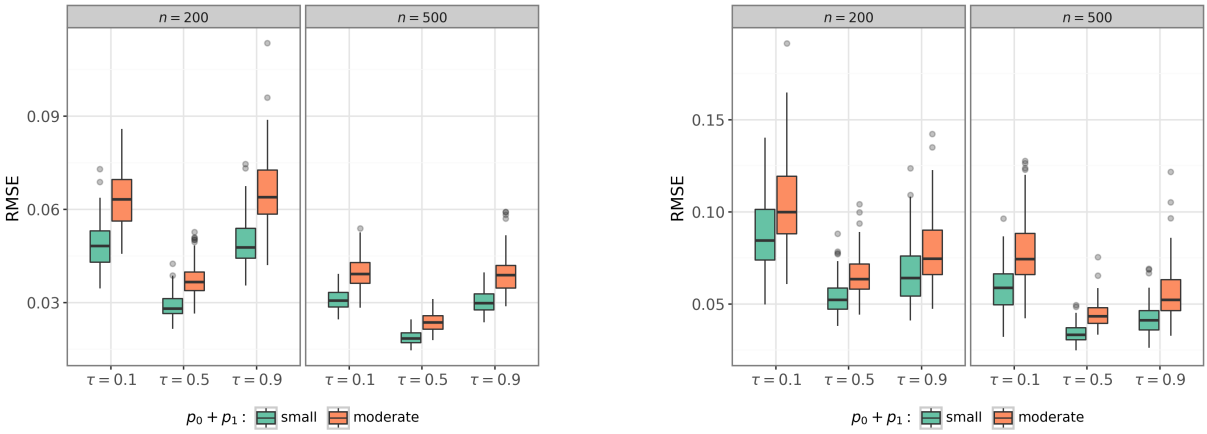


图2：在情景 S1（左）和 S2（右）下，所有后验样本中预测分位数的 RMSE 箱线图。

最后，我们计算以下函数95% 可信区间的覆盖率：$Q_{y_i|y_i\in(0,\ 1)}(\tau|\boldsymbol{x}_i)$、 p_{0i}和 p_{1i} 。对于每个后验样本，我们计算这些函数在 100 个测试观测值上的预测值。对于每个重复样本和每个测试观测值，我们根据后验样本计算 95% 可信区间，并记录该区间是否覆盖真实值（若是则为 1，若否则为 0）。首先，我们对所有重复样本中的这些指标取平均值，得到每个测试观测值的覆盖率；然后再对所有测试观测值取平均值，得到总体覆盖率。结果见图3。

总体而言，在情景 S1 和 S2 中， p_{0i} 和 p_{1i} 的覆盖率接近 95% 的名义水平；样本量越大，观察到 $\{0, 1\}$的概率越高，效果越好。在情景 S1 中，对于所有分位数，$Q_{y_i|y_i\in(0,1)}(\tau|\boldsymbol{x}_i)$ 的覆盖率也接近名义水平；但在情景 S2 中，对极端分位数（$\tau = 0.1$ 和 $\tau = 0.9$）进行建模时，覆盖率明显低于名义水平。除这一情况外，不同分位数的覆盖率相当一致。

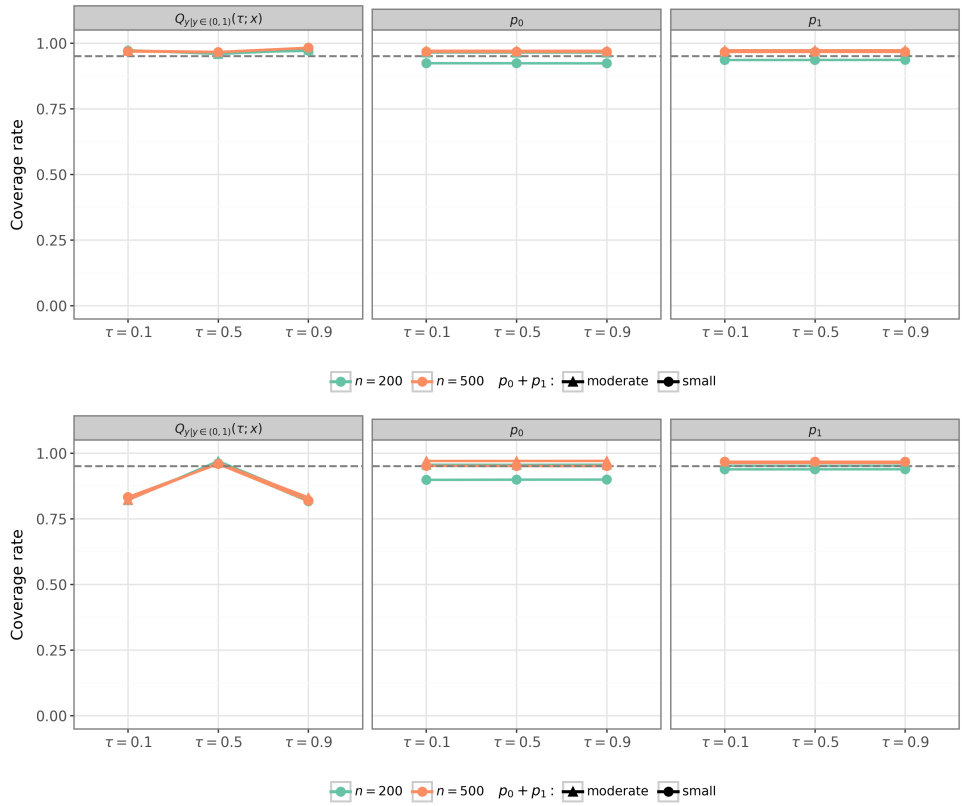


图3：在情景 S1（上排）和 S2（下排）中，τ分位数、p_{0i}和 p_{1i} 的95%置信区间覆盖率。灰色虚线表示名义上的95%覆盖水平。

5 应用

现在，我们通过两个实证应用来展示所提出模型的用法。这些示例旨在说明该建模框架的灵活性及其在不同情境下的表现。对于这两个应用，我们均运行四条 MCMC 链，每条链包含 1,500 次预热迭代和 5,000 次后验迭代。不进行抽稀，收集后验样本，从而共获得20,000 个样本用于推断。诊断结果表明，链的混合良好，且没有收敛问题的迹象。用于复现部分结果的代码可在 GitHub 仓库 inflated quantreg中获取。

5.1 交通事故死亡数据

我们使用了 $n = 5$个，499 巴西市辖区2002年交通事故死亡数据。数据来源于 DATASUS 数据库（https://datasus.saude.gov.br/）和 IPEADATA 数据库（https://www.ipeadata.gov.br/）。响应变量（y）是交通事故导致的死亡人数占比；协变量包括市镇居民人数的对数（`lnpop`）、年龄在20至29岁之间的居民比例（`prop2029`）、市辖区的教育人类发展指数（`hdie`）以及市辖区（`UF`）。Ospina 和 Ferrari（2012年）曾使用具有线性预测子的零膨胀贝塔回归模型分析这组数据的一个子集。在他们的分析中，未纳入协变量 `UF` 。本研究的主要目标是考察年轻人口（`prop2029`）对交通事故死亡人数占比的影响，同时控制潜在的混杂变量以及由协变量`UF`所刻画的空间结构。

响应变量呈现出明显的右偏分布，约有 38% 的市辖区报告交通事故死亡人数为零。尽管比例中位数为 0.024，均值却达到 0.048，这表明死亡率相对较高的市镇对均值产生了影响。观测到的最大值

为 0.80，出现在圣保罗州的一个市辖区。~图4 展示了各市辖区 y 的样本分位数。图4 显示，所有州 y 的第 10 百分位数都接近于零，表明许多市镇报告的交通事故死亡人数极少或为零。 y 在大多数州的中位数约为0.05；罗赖马州的中位数略高，约为0.13，其州界以绿色标出。部分州的第 90 百分位数接近0.3 ，包括罗赖马州、阿克里州（以蓝色标出）以及中西部地区的若干州，这表明这些地区的某些市镇死亡率较高。

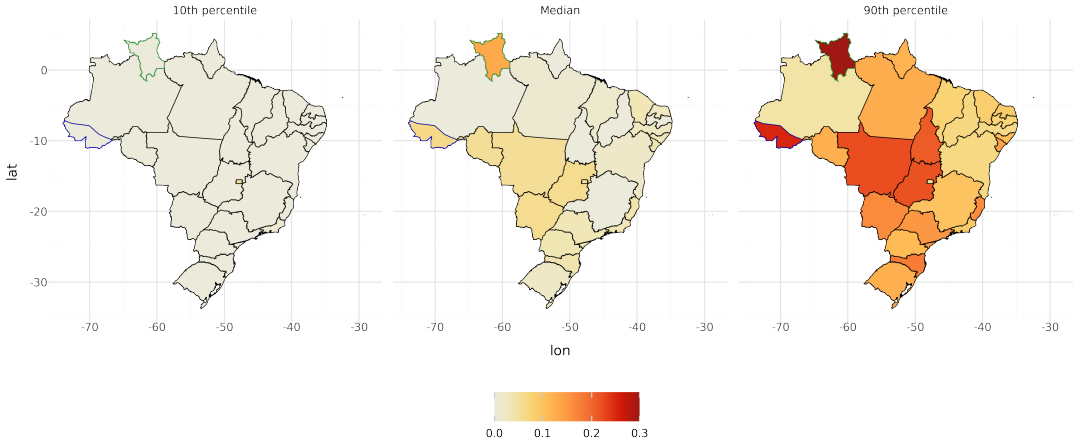


图4： y 在巴西各州（UF）的样本分位数空间分布：第10百分位数（$\tau = 0.1$，左图）、中位数（$\tau = 0.5$，中图）和第90百分位数（$\tau = 0.9$，右图）。

我们使用以下设定的零膨胀结构加性分位数回归对这些数据建模

$$\begin{aligned} \text{logit}(p_i) &= \beta_{10} + f_{11}(\texttt{lnpop}) + f_{12}(\texttt{prop2029}) + f_{13}(\texttt{hdie}) + f_{1,\text{spatial}}(\texttt{UF}), \\ \text{logit}\left[Q_{y_i|y_i\in(0,1)}(\tau \mid \mathbf{x}_i)\right] &= \beta_{20} + f_{21}(\texttt{lnpop}) + f_{22}(\texttt{prop2029}) + f_{23}(\texttt{hdie}) + f_{2,\text{spatial}}(\texttt{UF}), \end{aligned} \tag{10}$$

其中， $p_i = \mathbb{P}(y_i = 0)$是观测值为零的概率， $Q_{y_i|y_i\in(0,1)}(\tau \mid \mathbf{x}_i)$是给定 $y_i \in (0,1)$时 y_i 条件分布的第 τ分位数， β_{10} 和 β_{20} 分别是截距； f_{j1}、 f_{j2}和 f_{j3}、 $j = 1,2$分别是协变量lnpop、prop2029和hdie的非线性效应，这些效应通过含有20个基函数的三次惩罚样条建模，并采用弱逆伽马先验控制平滑度（形状和尺度均为0.01）。空间效应$f_{j,\text{spatial}}$(UF)、 $j = 1,2$通过定义在巴西27个州上的高斯马尔可夫随机场建模。邻接结构依据各区域的空间邻接关系定义，因此相邻州被视为邻居。空间效应的平滑度由方差参数控制，并采用形状和尺度均为0.01的弱逆伽马先验。

对于连续部分，我们对 5%、10%、20%、30%、50%、70%、80%、90% 和 95% 的条件分位数进行建模。这一选择使我们不仅能够评估各市镇呈现的中心模式（由中位数体现），还能够考察在至少报告了一起交通事故死亡的市镇中，交通事故死亡人数占比最小和最大的市镇的情况。此外，零膨胀结构明确对观察到零死亡的概率进行建模，使模型能够纳入没有交通事故死亡的市镇。

图5 展示了连续部分非线性效应估计值的图，以及空间效应估计值的图。这里，我们展示了六个选定分位数的结果；其余分位数的结果见补充材料第 2 节，因为它们与 10% 分位数的结果十分相似。请注意，由于连续部分仅对取值在 $(0,\ 1)$内的观测值建模，因此这些效应应理解为以响应变量取值落在该区间内为条件，即不包括死亡人数为零的市镇。换言之，所建模的分位数仅涉及至少观测到一例交通事故死亡的市镇。

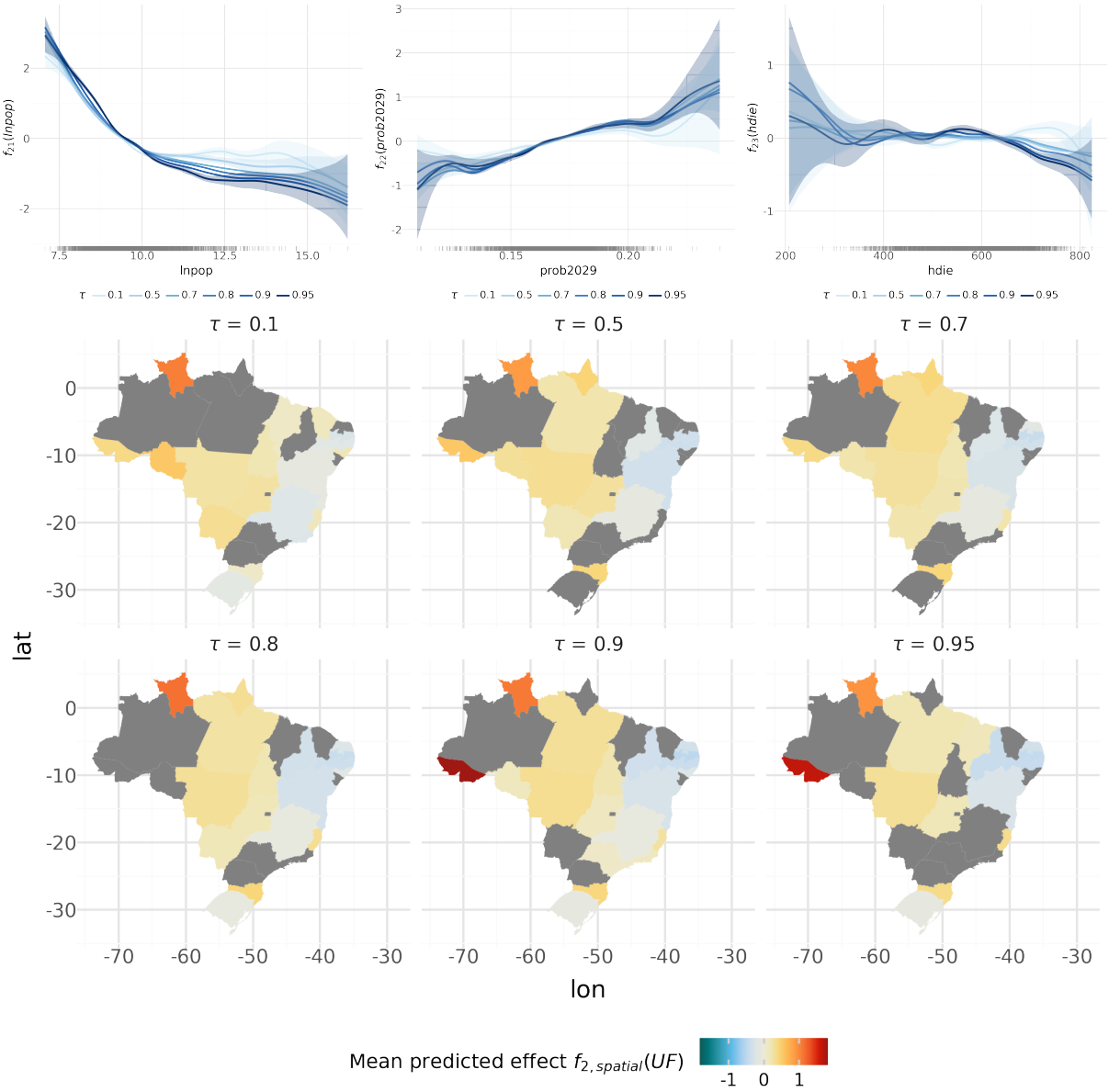


图5：连续部分估计效应的图。顶行展示了估计的非线性效应及其 90% 可信区间。为便于可视化，仅绘制了 10%、50% 和 90% 分位数的可信区间。中间行和底行展示了不同分位数下估计的空间效应。空间效应的 90% 可信区间包含零（不显著）的州以灰色显示。

lnpop 的效应估计表明，在所有条件分位数上，市镇规模与交通事故死亡人数占比总体呈负相关。总体而言，人口较多的市镇往往交通相关死亡率占比较低。人口规模较小时，这种下降最为明显，表明从极小型市镇发展到中等规模市镇，与交通死亡率相对大幅下降相关。人口规模较高时，这种效应减弱，说明随着人口规模增加，变化幅度逐渐减小。尽管这种关联在各分位数上的方向一致，但其大小有所不同。具体而言，与中位数或较低分位数相比，人口规模在上分位数（80%、90% 和 95%）上的负效应似乎更为明显，这表明人口规模与交通事故死亡占比较高的市镇交通死亡率降低相关。实际上，较大的市镇似乎更能缓解较高水平的交通事故死亡，这可能反映了基础设施、交通执法和医疗保健可及性方面的差异。

在所有分位数上，20–29岁居民所占比例（prop2029）与交通相关死亡所占比例之间均呈正相关。年轻居民占比较高的市镇，在发生死亡事故时往往交通死亡率也较高。这种关系大致呈线性，但可粗略分为三个阶段：第一和第三阶段的增幅略快，而中间阶段的增幅稍慢，整体趋势均向上。这一模式表明，在年轻居民占比较低的市镇，即使年轻人口比例略有增加，也与交通死亡率明显上升相关。在年轻居民占比适中的市镇，进一步增加的影响较小；而在占比较高的市镇，继续增加则会再次导致交通事故死亡人数明显上升。

在约200至630的较大取值范围内，所有分位数上hdie的估计效应都非常接近于零，这表明在教育发展水平较低至中等的市镇，教育发展指数与交通相关死亡所占比例之间没有明显关联。超出这一范围后，不同分位数上的效应差异变得更大。对于较低分位数（$\tau = 0.1$），hdie取值较高时效应略有上升，随后在最高水平时下降。相比之下，对于中位数和上分位数（$\tau \geq 0.5$），hdie超过约630后效应转为负值，且分位数越高，下降越明显，尤其是在 $\tau = 0.9$ 和 $\tau = 0.95$处。这一模式表明，在交通死亡率最高的市镇中，教育发展水平越高，交通相关死亡所占比例越低。

估计的空间效应揭示了巴西各地交通死亡率的区域模式。位于北部地区的阿克里州和罗赖马州呈现出持续为正的空间效应，尤其是在较高分位数处，这表明在至少发生一起死亡事故的条件下，这些州的市镇交通相关死亡占比相对较高。相比之下，圣保罗州是巴西人口最多、城市化程度最高的州，其各分位数的效应均接近于零。除这些特定情况外，空间效应在不同地区间有所差异：北部地区总体呈现正效应，东北部的效应往往为负或存在异质性，尤其是在较高分位数处；中西部的效应大多接近于零；南部地区则呈现较小的负效应或接近于零的效应。总体而言，较高分位数处的空间效应幅度更为显著，表明交通死亡率相对较高的市镇中，区域异质性更强。

图 6 展示了离散部分的估计效应。 `lnpop` 的估计效应呈下降趋势，且近似线性。这表明人口较多的市镇报告交通事故零死亡的可能性较低。 `prop2029` 的估计效应在大多数取值范围内接近于零，仅在 0.13 至 0.15 之间出现一个小幅正峰值，这表明年轻人口占比处于该范围的市镇，报告交通事故零死亡的概率可能略高。 `hdie`的估计效应呈下降趋势。对于教育发展水平较低的市镇（`hdie` 低于 500），效应略为正向，表明报告零死亡的概率较高。对于教育发展水平中等至较高的市镇（`hdie` 高于 500），效应转为略微负向，但仍非常接近于零，表明其影响极小。关于交通事故零死亡概率的空间效应，北部和东北部地区的大多数州呈现正效应，说明这些地区的市镇报告零死亡的可能性较高。南部、东南部及中西部部分地区则呈现负效应，表明零死亡概率较低。这些结果表明，北部和东北部某些州的市镇交通事故死亡人数往往较少，而城市化程度较高、人口密集的地区则更可能记录至少一起死亡事故。

总体而言，综合连续部分和零膨胀部分的结果表明，尽管较大的市镇往往交通事故死亡更为常见，但其在总体死亡率中所占比例较小。年轻居民比例对零值概率影响甚微，但在连续部分中与交通死亡率呈正相关，这表明人口较年轻的市镇在发生死亡事故时，交通相关死亡所占比例往往更高。教育发展指数与零值概率的关联较弱，但在较高分位数上与交通相关死亡比例呈负相关，这表明教育发展水平越高，非零交通相关死亡所占比例越小。最后，空间效应揭示了显著的区域差异。

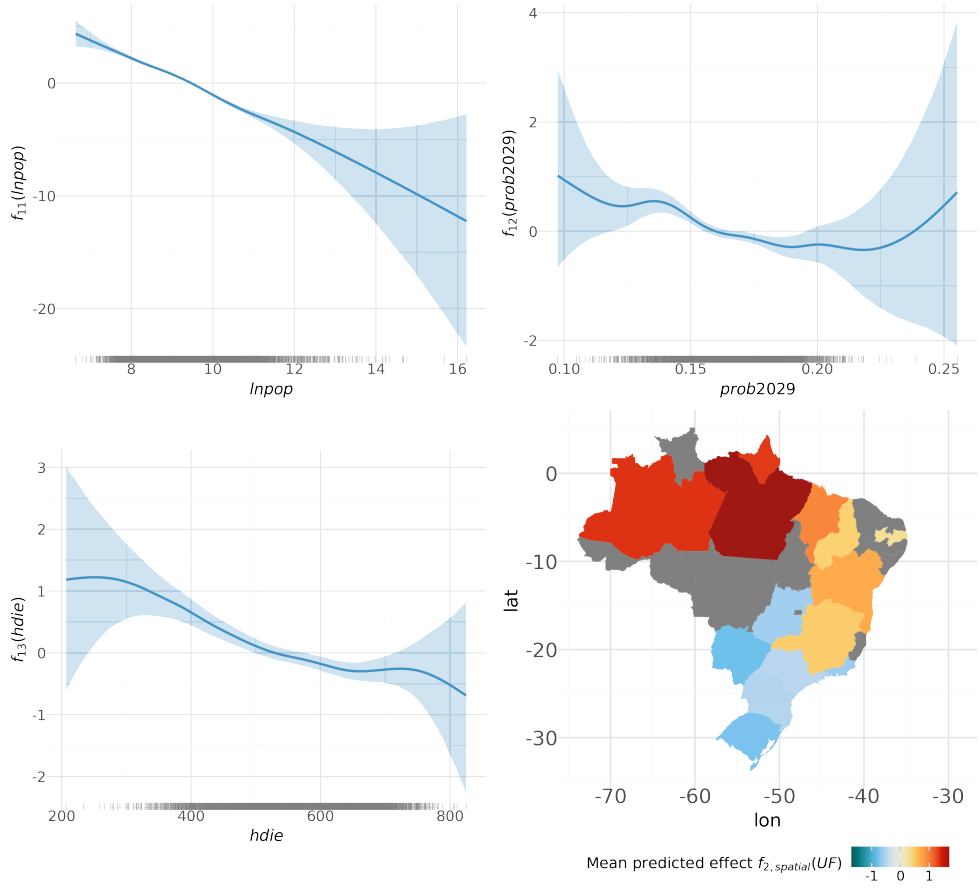


图6：离散部分估计效应图。右下图显示估计的空间效应，其余各图显示估计的非线性效应。空间效应的 90%可信区间包含零（不显著）的州以灰色表示。

北部和东北部州的市镇报告零死亡事故的概率往往较高，但一旦发生死亡事故，其中一些州，尤其是北部地区的州，条件死亡率水平则较高。相反，南部和东南部等城市化程度较高的地区，零值概率较低，但非零交通死亡率水平通常也较低。

为了进行比较，我们还对这些数据拟合了零膨胀贝塔回归模型（Ospina & Ferrari，2012年）。我们考虑条件均值（μ_i）、条件精度（ϕ_i）和零值概率（p_i）预测变量的以下设定

$$
\begin{aligned}
\text{logit}(p_i) &= \beta_{10} + f_{11}(\texttt{lnpop}) + f_{12}(\texttt{prop2029}) + f_{13}(\texttt{hdie}) + f_{1,\text{spatial}}(\texttt{UF}),\\
\text{logit}(\mu_i) &= \beta_{20,\mu} + f_{21,\mu}(\texttt{lnpop}) + f_{22,\mu}(\texttt{prop2029}) + f_{23,\mu}(\texttt{hdie}) + f_{2,\mu,\text{spatial}}(\texttt{UF}),\\
\log(\phi_i) &= \beta_{20,\phi} + f_{21,\phi}(\texttt{lnpop}) + f_{22,\phi}(\texttt{prop2029}) + f_{23,\phi}(\texttt{hdie}) + f_{2,\phi,\text{spatial}}(\texttt{UF}).
\end{aligned}
$$

我们使用 Liesel 拟合该模型，并保留与零膨胀结构加性分位数回归模型方程（10）相同的函数形式、基函数数量、空间效应和回归参数先验设定。

由于两个模型中的离散成分相同，我们仅呈现连续分量的结果。图 7 展示了条件均值和精度子模型的估计非线性效应与空间效应。零膨胀贝塔回归中条件均值的估计效应与零膨胀结构加性分位数回归在中位数分位数处得到的结果非常相似。对于条件精度参数，`lnpop` 的估计效应近似为线性且为正，表明较大的市镇往往具有更高的精度。`prop2029` 的效应略有下降，表明

随着年轻居民占比的增加，精度略有下降。对于 hdie，该效应在中等取值附近接近于零，但在较高取值处呈现凸形模式，表明教育发展水平较高的市镇中条件精度有适度上升。对条件精度子模型中空间效应的分析表明，总体而言，负效应出现在南部、东南部和北部地区的部分区域，表明这些地区的精度较低、变异性较大。正效应主要集中在东北部的大部分地区，表明精度较高、变异性较低。值得注意的是，阿克里州表现出强烈的负效应，凸显该州的市镇经历了更多变的交通死亡模式。

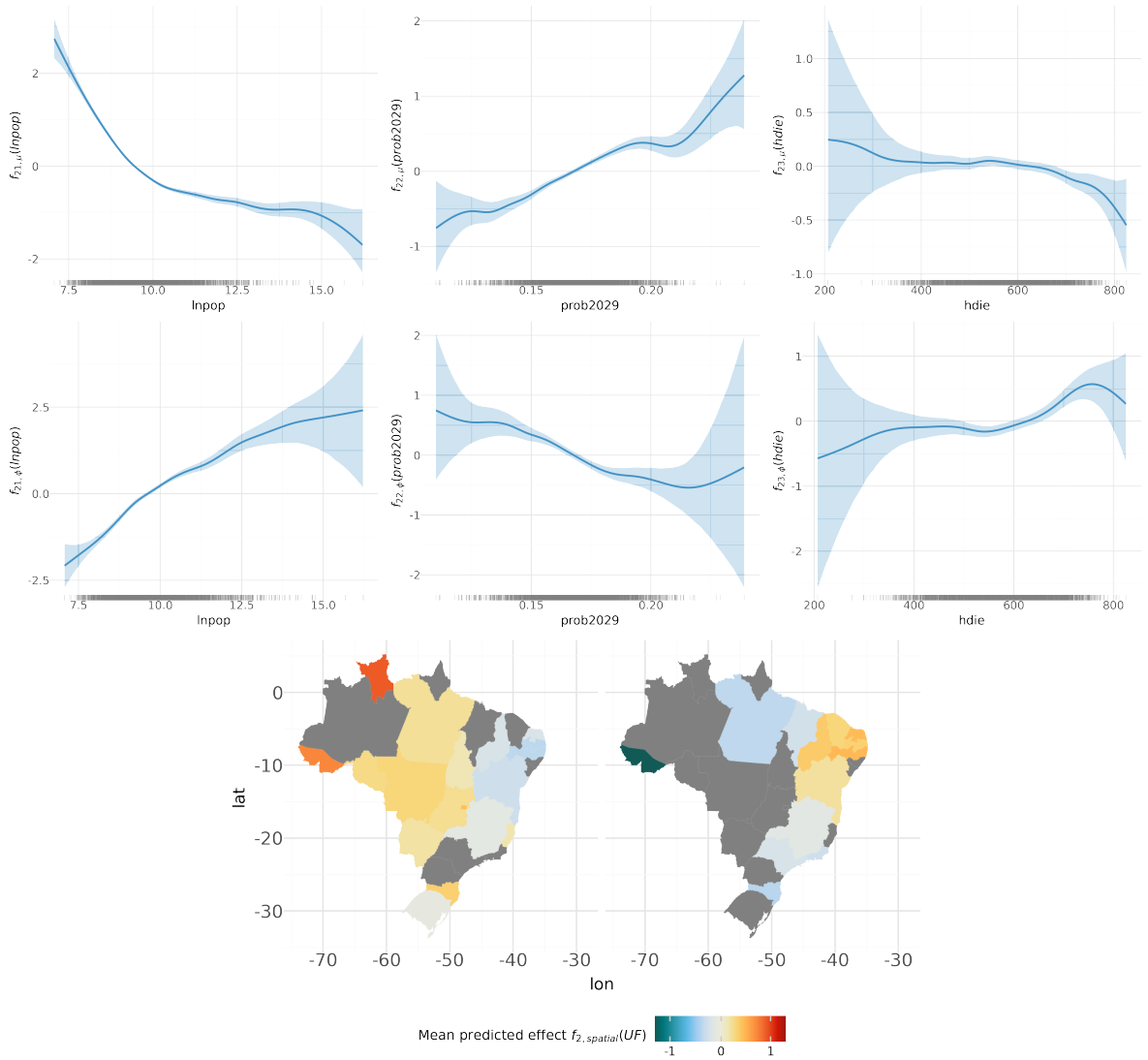


图7：零膨胀贝塔回归模型连续部分的效应估计图。顶行显示条件均值子模型的非线性效应估计，中间行显示条件精度子模型的非线性效应估计，底行显示空间效应估计。空间效应的90%可信区间包含零（不显著）的州以灰色显示。

最后，图8 展示了10%、50%、70%、80%、90%和95% 分位数在零膨胀结构加性分位数回归和零膨胀贝塔回归模型下的拟合条件分位数。在所有分位数以及两种模型中，拟合分位数均随lnpop下降，随prop2029上升，并与hdie呈略微下降的关系。较低分位数（$\tau = 0.1$）通常非常接近于零，并且在两种建模方法之间几乎无法区分。这一模式在所有州中均可观察到。对于S˜ao Paulo和罗赖马州，拟合分位数在零膨胀结构加性分位数回归与零膨胀贝塔回归模型之间非常相似，两种模型的估计几乎没有可见差异，在所有分位数

水平上。相比之下，对于阿克里州，拟合分位数在不同模型之间的差异更为明显。特别是，对于中间分位数（$\tau = 0.5$、0.7和0.8），零膨胀贝塔回归往往产生系统性地高于分位数回归的拟合分位数，表现为估计曲线的上移。这种差异在上分位数（$\tau = 0.9$ 和0.95）处更加明显，两种模型的拟合曲线差异更为清晰。在这些水平上，分位数回归得到的估计行为更为一致，可信区间也相对更为集中，而贝塔回归在上尾表现出更大的不确定性，尤其是对于prop2029和hdie。

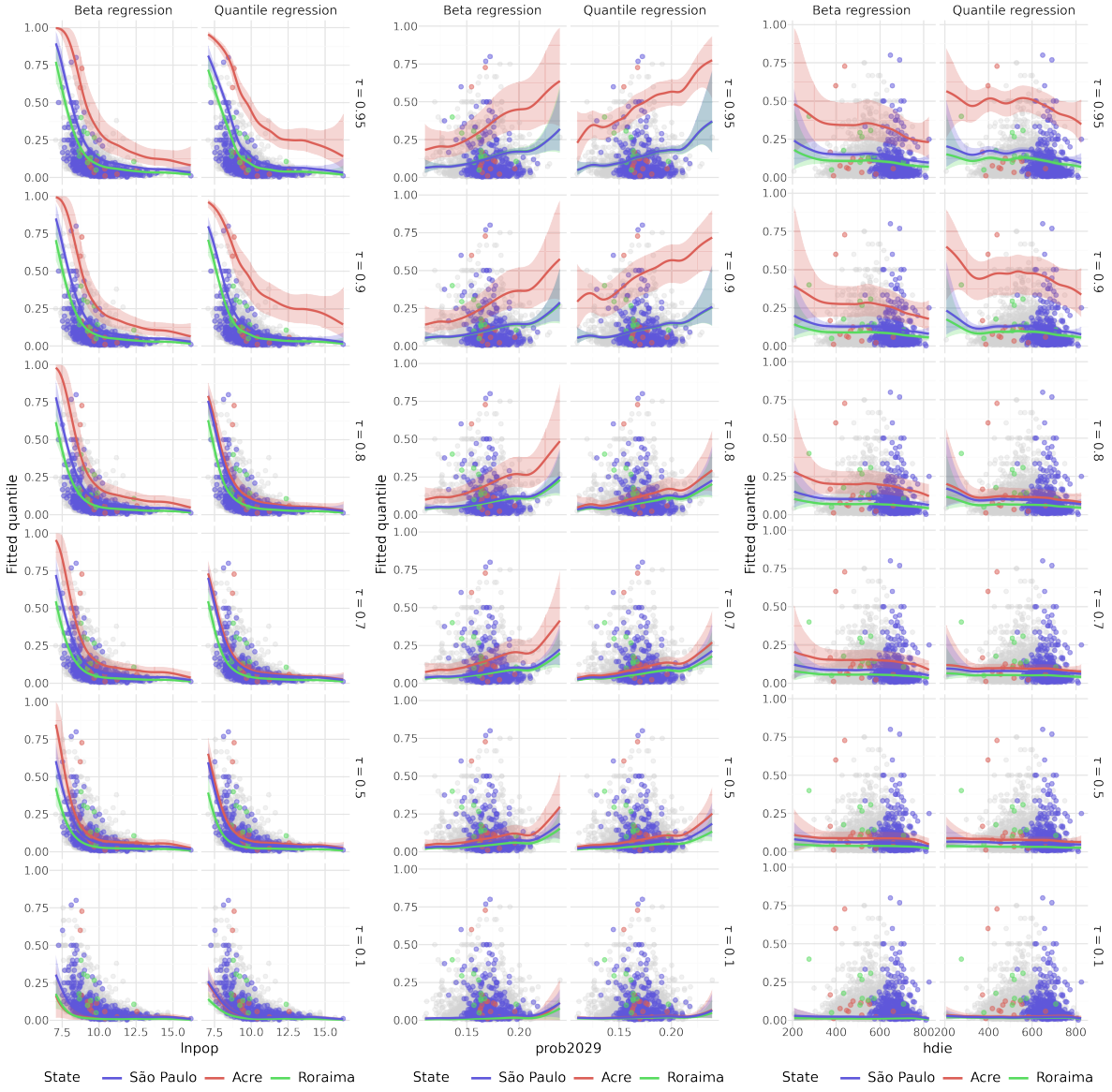


图8：基于零膨胀结构加性分位数回归和零膨胀贝塔回归模型得到的连续部分拟合分位数。实线表示阿克里州、罗赖马州和圣保罗州预测的 τ 分位数的后验均值。所有其他协变量均固定为各州样本均值。阴影带表示 90% 可信区间。

总体而言，零膨胀贝塔回归和零膨胀结构加性分位数回归对协变量效应如何影响交通死亡比例的方向得出了大体一致的结论。不过，分位数回归进一步揭示了这些效应在条件分布不同部分之间的变化情况。具体而言，人口规模与交通死亡比例之间的负相关，以及年轻居民比例与交通死亡比例之间的正相关，在较高

分位数处更为显著，表明在人口交通死亡比例相对较高的市镇中，这些效应更强。贝塔回归侧重于条件均值，无法直接捕捉这些分布差异。零膨胀贝塔回归仍是一种实用工具，而零膨胀分位数回归则能更详细地刻画协变量效应在交通死亡比例整个条件分布中的变化。

5.2 言语可懂度测试数据

我们分析了 Hu 等人（2015年）在一项语音可懂度实验中呈现的数据。该实验涉及七名人工耳蜗植入者，旨在评估不同声音处理算法下语音识别表现的差异。这些数据也由 Stasinopoulos 等人（2024年，第9章）进行了分析。在实验中，受试者聆听预先录制的句子，句子在背景噪声（干扰说话人）中播放；不同试次的信噪比（SNR）各不相同，受试者需要复述句子。数据按测试组收集，每组包含20个句子，同一测试组内的聆听条件保持不变。在本研究中，每个句子对应一条观测记录，响应变量为正确识别的语素比例，定义为 $y = w/N$，其中 w 表示正确识别的语素数量，N 表示句子中的语素总数。

该数据集包含以下协变量：人工耳蜗中使用的声音处理算法（algorithm：A、B 或 C）、干扰说话人的性别（noise g：男性或女性），以及受试者标识符（subject，可在分析中纳入受试者层面的随机效应）。每名受试者在每种实验条件下完成四个测试组，因此每种条件包含80个句子。

这些数据的一个显著特征是极端响应出现的频率很高：许多句子要么被完全识别（$y = 1$），要么完全未被识别（$y = 0$），每种极端情况约占句子的34%。因此，几乎68% 的观测值位于单位区间的边界处 $[0, 1]$。Hu 等人（2015年）使用零膨胀和一膨胀贝塔回归模型分析了这些数据，而 Stasinopoulos 等人（2024年，第9章）则使用了零膨胀和一膨胀单纯形回归模型。在本研究中，我们采用零膨胀和一膨胀结构化加性分位数回归模型，并对预测变量作如下设定

$$
\begin{aligned}
\log\left(\frac{p_{0ij}}{p_{2ij}}\right) &= \beta_{00} + \beta_{11}\mathtt{snr}_{ij} + \beta_{02}\mathtt{algorithmB}_{ij} + \beta_{03}\mathtt{algorithmC}_{ij} + \beta_{04}\mathtt{noise_gM}_{ij} \\
&\quad + \beta_{05}\mathtt{snr}_{ij} \times \mathtt{algorithmB}_{ij} + \beta_{06}\mathtt{snr}_{ij} \times \mathtt{algorithmC}_{ij} + u_{0j}, \\
\log\left(\frac{p_{1ij}}{p_{2ij}}\right) &= \beta_{10} + \beta_{11}\mathtt{snr}_{ij} + \beta_{12}\mathtt{algorithmB}_{ij} + \beta_{13}\mathtt{algorithmC}_{ij} + \beta_{14}\mathtt{noise_gM}_{ij} \\
&\quad + \beta_{15}\mathtt{snr}_{ij} \times \mathtt{algorithmB}_{ij} + \beta_{16}\mathtt{snr}_{ij} \times \mathtt{algorithmC}_{ij} + u_{1j}, \\
\operatorname{logit}\left[Q_{y_{ij}|y_{ij}\in(0,1)}(\tau \mid \mathbf{x}_{ij})\right] &= \beta_{20} + \beta_{21}\mathtt{snr}_{ij} + \beta_{22}\mathtt{algorithmB}_{ij} + \beta_{23}\mathtt{algorithmC}_{ij} + \beta_{24}\mathtt{noise_gM}_{ij} \\
&\quad + \beta_{25}\mathtt{snr}_{ij} \times \mathtt{algorithmB}_{ij} + \beta_{26}\mathtt{snr}_{ij} \times \mathtt{algorithmC}_{ij} + u_{2j},
\end{aligned}
$$

其中，$p_{0ij} = \mathbb{P}(y_{ij} = 0)$ 和 $p_{1ij} = \mathbb{P}(y_{ij} = 1)$ 分别表示观测到零和一的概率，$p_{2ij} = 1 - p_{0ij} - p_{1ij}$，$Q_{y_{ij}|y_{ij}\in(0,1)}(\tau \mid \mathbf{x}ij)$ 表示在给定 y_{ij} 且 $y_{ij} \in (0,1)$ 的条件下，条件分布的第 τ 个分位数，其中对应第 i 个句子和第 j 个受试者；β_{00}、β_{01} 和 β_{02} 为截距，其余系数对应于协变量 snr、算法B、算法C、噪声 gM 及其交互项的线性效应。预测变量的设定基于 Stasinopoulos 等人（2024年，第9章）开展的变量选择。受试者间变异通过随机截距 u_{0j}、u_{1j}、u_{2j} 加以刻画，这些随机截距建模为相互独立的高斯项，其方差参数服从浓度为 1、尺度为 0.005 的弱逆伽马先验。我们考虑三个条件分位数（10%、50%、90%），以刻画识别得分的中心趋势和总体分布。此外，零膨胀和一膨胀结构还对相对于连续响应的完全未识别句子（$y = 0$）和完全正确句子（$y = 1$）的对数优势进行建模。

表1 给出了模型参数的后验均值和标准差，以及相应的90% 可信区间。为便于建模，所有分类变量均编码为 $-1/1$ 虚拟变量。声音处理算法以算法B和算法C表示，其中1 表示存在相应算法，而 -1 表示参考算法（A）。类似地，噪声 gM

编码为1 表示男性说话者， −1表示女性说话者。在此编码方案下，每个系数反映从一个类别变为另一个类别时总效应的一半。表1还报告了受试者层级随机截距方差的后验均值和标准差。离散部分子模型的结果与 Stasinopoulos 等人（2024年，第9章）的结果相似；连续部分在中位数处的结果（$\tau = 0.5$）也与其条件均值子模型的结果相当，细微差异可能源于稳健性。

表1：三个子模型的后验分布汇总： $\tau = 0.1$，0。5、0。9处的分位数子模型，以及离散子模型 log(p_{0ij}/p_{2ij}) 和 log(p_{1ij}/p_{2ij})。对于每个参数，报告后验均值、标准差和90% 可信区间。最后一行报告受试者层级随机截距方差的后验均值和标准差。

	$\text{logit}[Q_{y_{ij}\mid y_{ij}\in(0,1)}(\tau \mid \mathbf{x}^{ij})]$								
	$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
	Mean	SD	90%置信区间	Mean	SD	90%置信区间	Mean	SD	90%置信区间
截距	−1.646	0.040	(−1.713, −1.583)	−0.308	0.060	(−0.407, −0.211)	1.505	0.045	(1.428, 1.576)
snr	0.075	0.010	(0.059, 0.091)	0.095	0.011	(0.076, 0.113)	0.027	0.008	(0.014, 0.040)
算法B	0.026	0.035	(−0.032, 0.082)	0.067	0.057	(−0.025, 0.161)	0.022	0.035	(−0.036,0.080)
算法C	0.011	0.033	(−0.043, 0.064)	0.017	0.050	(−0.064, 0.098)	−0.017	0.042	(−0.088, 0.052)
噪声 gM	0.041	0.022	(0.005, 0.076)	0.051	0.026	(0.009, 0.094)	0.033	0.015	(0.008, 0.059)
信噪比：算法B	0.009	0.010	(−0.007,0.026)	−0.011	0.013	(−0.033, 0.010)	0.001	0.008	(−0.012,0.013)
信噪比:算法C	−0.011	0.010	(−0.027,0.005)	−0.022	0.012	(−0.042, −0.003)	−0.001	0.008	(−0.014,0.012)
随机（受试者）	0.003	0.003		0.007	0.007		0.003	0.002	

	$\log(p_{0ij}/p_{2ij})$			$\log(p_{1ij}/p_{2ij})$		
	Mean	SD	90%置信区间	Mean	SD	90%置信区间
截距	0.633	0.089	(0.488,0.776)	−0.909	0.143	(−1.140, −0.677)
snr	−0.221	0.015	(−0.245, −0.197)	0.213	0.014	(0.190, 0.237)
算法B	−0.130	0.058	(−0.226, −0.034)	0.060	0.073	(−0.060, 0.181)
算法C	−0.174	0.060	(−0.273, −0.074)	−0.131	0.075	(−0.254, −0.007)
噪声 gM	−0.161	0.032	(−0.214, −0.108)	0.151	0.033	(0.098, 0.205)
信噪比：算法B	0.050	0.016	(0.024,0.075)	−0.027	0.015	(−0.051, −0.002)
信噪比:算法C	0.011	0.016	(−0.016,0.038)	−0.014	0.015	(−0.039, 0.011)
随机（受试者）	0.036	0.030		0.118	0.088	

条件分位数结果展示了实验条件与人工耳蜗使用者不同水平的语音识别表现之间的关联，重点关注部分理解的句子，即排除完全正确和完全未识别的回答。根据表1，在所有分位数上，较高的信噪比与部分识别语素的条件分位数值较高相关；这种效应在较低至中位数分位数处（$\tau = 0.1$ 和 0.5）更强，在较高分位数处（$\tau = 0.9$）则较弱。干扰说话人的性别在所有分位数上也呈现出较小但为正向的效应。此外，在 $\tau = 0.5$处观察到信噪比与算法 C 之间存在负向交互作用。对于离散部分，较高的信噪比与句子完全未识别的优势降低、句子完全正确的优势提高相关，这表明更好的聆听条件与更倾向于部分或完全识别相关。最后，男性干扰说话人与部分或完全识别的优势较高相关，同时观察到信噪比与算法 C 之间存在交互作用。

图9展示了总结连续部分和离散部分受试者水平随机截距的森林图。这些图揭示了受试者间变异的不同模式。在连续部分中，估计的随机效应总体较小，后验均值接近零，且90% 可信区间大多包含零，表明受试者间变异有限。尽管如此，它们对于考虑每位受试者的重复测量仍然重要。相比之下，离散分量中的受试者间变异更大，这体现在受试者水平随机截距的后验方差中；另见表1。图9表明，对于离散部分的子模型，若干受试者的可信区间并不包含

零，表明他们产生极端结果的倾向存在差异。总体而言，这突出表明个体差异在完全失败或完美识别的概率上最为显著，而中间表现在各受试者之间相对同质。

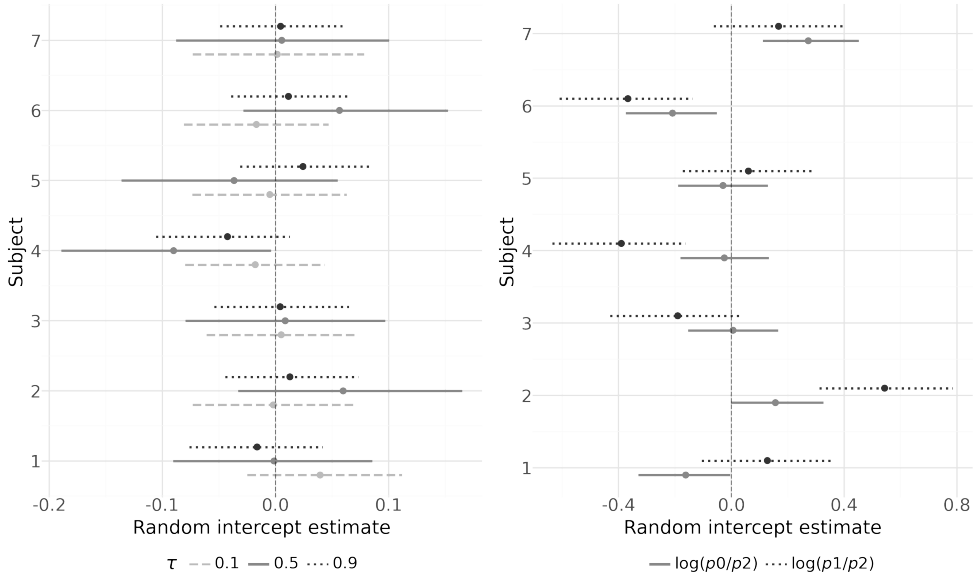


图 9：汇总各受试者随机截距的森林图。左侧面板显示连续部分在 0.1、0.5 和 0.9 分位数处的估计值，右侧面板显示离散子模型的估计值。每条水平线表示某位受试者随机截距的后验均值，线段长度表示其 90% 可信区间。

为进一步考察三种声音处理算法在表现上的差异，图 10展示了连续部分拟合的条件分位数，以及完全失败（$y = 0$）和完美识别（$y = 1$）的拟合概率随信噪比变化的情况，分别针对女性和男性干扰说话者。为便于可视化，将受试者水平随机截距设为零。总体而言，我们观察到，随着信噪比增加，正确识别的语素比例也趋于上升。此外，拟合的分位数曲线和概率曲线显示，女性与男性干扰说话者之间差异很小。

对于较高分位数（$\tau = 0.9$），三种算法在整个信噪比范围内的表现非常相似。这表明，对于那些既未完全漏听、也未被完全识别的句子中识别比例相对较高的句子，算法的选择几乎没有影响。对于 $\tau = 0.5$，即连续部分的条件中位数，低信噪比下算法 B 和 C 的拟合值高于算法 A，表明它们可能有助于在更具挑战性的聆听条件下进行识别。在信噪比的中间范围内，所有算法的表现似乎相近，说明在中等聆听条件下，算法的影响相对较小。在高信噪比下，算法 A 和 B 往往优于算法 C，这表明在更有利的聆听条件下，算法 C 的效果可能略逊一筹。关于条件中位数分位数的这些结果，与 Stasinopoulos 等人（2024年，第9章）针对零膨胀和一膨胀单纯形回归模型条件均值所报告的结果一致。对于较低分位数（$\tau = 0.1$），即那些既未完全漏听、也未被完全识别的句子中识别比例最低的句子，在低至中等信噪比下，三种算法的表现相似。从中等到高信噪比，算法 B 得到的拟合值最高，算法 A 居中，算法 C 最低。这表明，对于更难识别的句子，算法 B 可能带来最大的益处；而即使在相对有利的聆听条件下，算法 C 的效果似乎仍然较差。

最后，根据图10，完全漏听句子的拟合概率（$y = 0$）和完全正确句子的拟合概率（$y = 1$）显示，算法性能会因信噪比而异。对于零值的拟合概率（p_0），在低信噪比下，算法 A 对完全漏听句子的拟合概率最高，算法 C 居

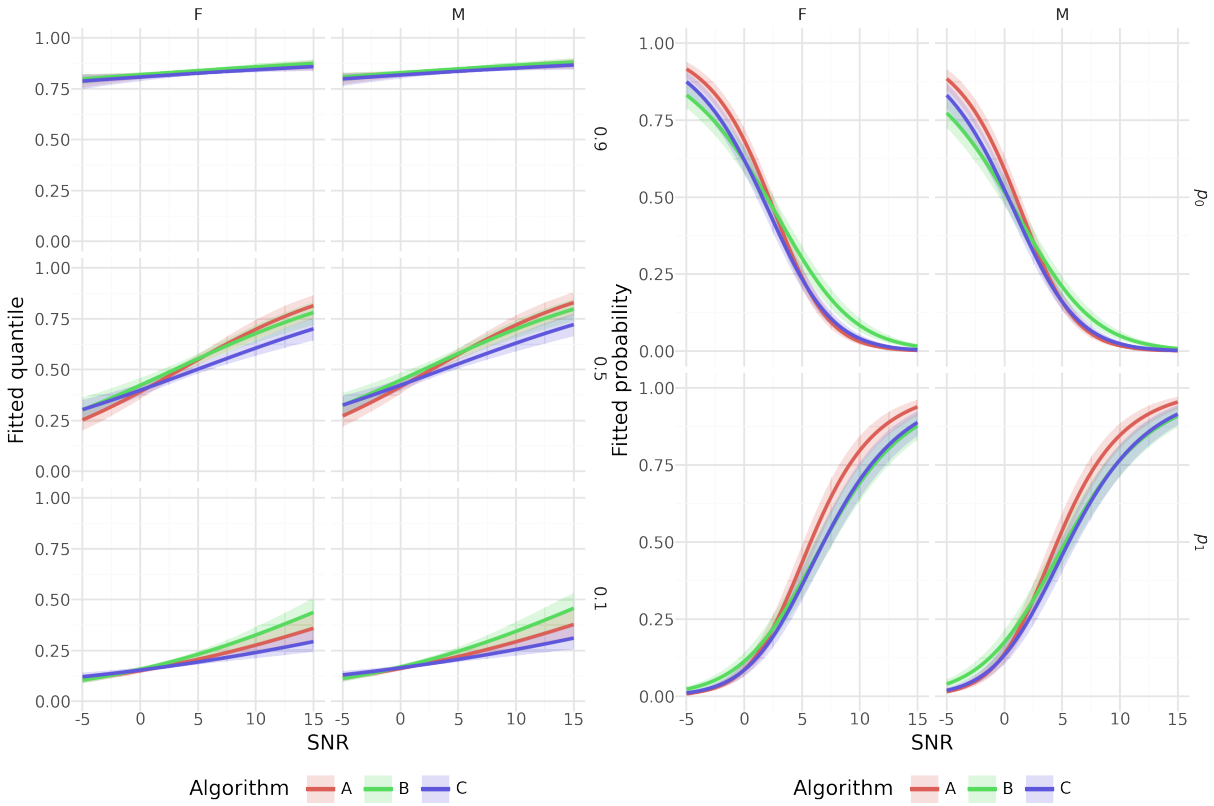


图10：针对女性（左列）和男性（右列）噪声性别，按算法A、B和C绘制的拟合分位数（左图）和概率（右图）随信噪比的变化，并给出90%可信区间。左图显示连续部分在 $\tau = 0.1$（底部）、0.5（中间）和0.9（顶部）处拟合的条件分位数；右图显示零值（p_0，顶部）和1（p_1，底部）的拟合概率。

中，算法 B 最低，这表明在具有挑战性的聆听条件下，算法 B 可能有助于减少完全失败。在信噪比中等时，所有算法的零值概率往往相近，表明它们之间差异很小。在高信噪比下，算法 A 和 C 的零值拟合概率相近，而算法 B 略高，这表明在有利条件下，算法 B 的失败次数可能略多。对于一值的拟合概率（p_1），在低信噪比下，所有算法的拟合概率相近，算法 B 略高，这表明在聆听条件困难时，各算法之间仅有细微差异。在信噪比中等时，所有算法表现相似。从中等到高信噪比，算法 A 的完全正确响应拟合概率最高，算法 B 和 C 非常相近，这表明在良好的聆听条件下，算法 A 可能更有助于实现完美识别。总体而言，综合考虑零值和一值的概率，算法 B 在低信噪比下表现似乎更好，因为它能减少完全失败；而在高信噪比下，算法 A 似乎更有优势，其完全正确句子的概率略高。

6 结论

本文提出一种灵活的分位数回归模型，用于处理边界处存在观测值的有界连续数据。零值或一值膨胀有界数据的结构化加性分位数回归模型，在Santos和Bolfarine（2015年）提出的零值或一值膨胀有界数据分位数回归模型基础上，纳入了结构加性回归，从而能够包含空间效应、随机效应、变系数项等。我们还提出了零膨胀和一膨胀结构化加性分位数回归模型，用于分析单位区间两端均有不可忽略数量观测值的有界连续数据。我们采用两部分建模框架，分别对条件

连续响应变量的分位数以及观测到零值和／或一值的概率进行建模。在这两种模型中，响应变量的条件分位数以及零值和／或一值的概率均通过连接函数与预测变量相关联。我们在贝叶斯框架下，使用非对称拉普拉斯分布和高效的马尔可夫链蒙特卡罗算法进行推断，并在Liesel概率编程框架中实现。模拟结果表明，所提出的模型能够准确估计离散和连续成分的参数，并具有良好的预测性能和可信区间覆盖率。

第5.1节展示了该模型在巴西市辖区交通相关死亡率数据中的应用，凸显了采用结构化加性预测变量充分捕捉复杂协变量效应的必要性。具体而言，结果表明，数据呈现明显的非线性和空间模式，而 Santos 和 Bolfarine（2015年）提出的零膨胀分位数回归模型仅限于简单的线性预测子，无法处理这些模式。第5.2节通过分析语音可懂度数据，展示了零膨胀和一膨胀结构化加性分位数回归模型的应用。在此应用中，所提出的模型能够纳入随机效应，以解释个体特异性变异，并恰当地捕捉数据的层级结构。

最后，这里提出的方法可以自然地扩展到对存在零膨胀的正值连续数据进行建模，即定义在 $[0, \infty)$ 上的数据。在我们的方法中，将条件分位数与加性预测变量关联起来的链接函数目前定义在 $(0,1)$上，因为连续响应被限制在单位区间内；参见公式（4）。对于零膨胀正值连续数据，可以通过将链接函数定义在 $(0, \infty)$上应用同一框架，从而使模型能够处理更广泛的有界或无界连续响应。

致谢

本研究部分由巴西圣保罗研究基金会（FAPESP）资助，项目编号为 2024年/08343-9。

参考文献

Bayes, C. L., Baz´an, J. L., & De Castro, M. (2017). 有界响应变量的分位数参数混合回归模型. Statistics and its Interface, 10(3), 483–493. https://doi.org/10.4310/SII.2017.v10.n3. a11 Bottai, M., Cai, B., & McKeown, R. E. (2010). 有界结果的逻辑分位数回归. Statistics in Medicine, 29, 309–317. https://doi.org/10.1002/sim.3781 Bourguignon, M., Gallardo, D. I., & Saulo, H. (2024). 参数分位数贝塔回归模型. International Statistical Review, 92(1). https://doi.org/10.1111/insr.12564 DeepMind, Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., . . . Viola, F. (2020). DeepMind JAX 生态系统 *[* Python库*]*. https: //github.com/google-deepmind Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2021). 回归：模型、方法与应用 (第2版). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-63882-8 Gamerman, D. (1997). 广义线性混合模型中的后验分布抽样. Statistics and Computing, 7, 57–68. https://doi.org/https://doi.org/10.1023/A:1018509429360 Hoffman, M. D., & Gelman, A. (2014). 无转向采样器：在哈密顿蒙特卡罗中自适应设定路径长度. Journal of Machine Learning Research, 15(1), 1593–1623. Hu, W., Swanson, B. A., & Heller, G. Z. (2015). 语音可懂度测试的统计分析方法. PLOS ONE, 10(7), 1–17. https://doi.org/10.1371/journal.pone.0132409

Klein, N., Kneib, T., Lang, S., & Sohn, A. (2015). 贝叶斯结构化加性分布回归及其在德国区域收入不平等研究中的应用. The Annals of Applied Statistics, 9(2), 1024–1052. https://doi.org/10.1214/15-AOAS823 Kneib, T. (2013). 超越均值回归. Statistical Modelling, 13(4), 275–303. https : / / doi . org / 10 . 1177/ 1471082X13494159 Kneib, T., Silbersdorff, A., & S¨afken, B. (2023). 对均值的反叛——分布回归方法综述. Econometrics and Statistics, 26, 99–123. https://doi.org/10.1016/j.ecosta.2021.07.006 Koenker, R. (2004). 纵向数据的分位数回归. Journal of Multivariate Analysis, 91(1), 74–89. https://doi.org/https://doi.org/10.1016/j.jmva.2004.05.006 Koenker, R., & Bassett, G. (1978). 回归分位数. Econometrica, 46(1), 33–50. https://doi.org/10.2307/ 1913643 Koenker, R. (2005). 分位数回归. 剑桥大学出版社. https://doi.org/https://doi.org/10.1017/ CBO9780511754098 Kosmidis, I., & Zeileis, A. (2025). 面向 $[0, 1]$响应的扩展支持域贝塔回归. Journal of the Royal Statistical Society Series C: Applied Statistics, 75(1), 139–157. https://doi.org/10.1093/jrsssc/qlaf039 Kozumi, H., & Kobayashi, G. (2011). 贝叶斯分位数回归的吉布斯抽样方法. Journal of Statis- tical Computation and Simulation, 81(11), 1565–1578. https://doi.org/10.1080/00949655.2010.496117 Lemonte, A. J., & Baz´an, J. (2016). 新型 Johnson S_B 分布及其率和比例回归模型. Biometrical Journal, 58(4), 727–746. https://doi.org/10.1002/bimj.201500030 Menezes, A. F., Mazucheli, J., & Bourguignon, M. (2021). 对零膨胀或一膨胀双界数据建模的参数分位数回归方法. Biometrical Journal, 63, 841–858. https://doi.org/10.1002/ bimj.202000126 Ospina, R., & Ferrari, S. L. P. (2010). 膨胀贝塔分布. Statistical Papers, 51, 111–126. https://doi.org/ 10.1007/s00362-008-0125-4 Ospina, R., & Ferrari, S. L. P. (2012). 一类通用的零膨胀或一膨胀贝塔回归模型. Computational Statistics & Data Analysis, 56(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005 Queiroz, F. F., & Ferrari, S. L. P. (2024). 热带金枪鱼迁移建模：一种膨胀幂次逻辑回归方法. Biometrical Journal, 66(4), 2300288. https://doi.org/https://doi.org/10.1002/bimj.202300288 Razen, A., Brunauer, W., Klein, N., Kneib, T., Lang, S., & Umlauf, N. (2023). 运用分布回归与分位数回归进行房地产估值的多层次分析. Statistical Modelling, 23(5-6), 525–539. https : // doi.org/https://doi.org/10.1177/1471082X23115720 Riebl, H., Wiemann, P. F. V., & Kneib, T. (2023). Liesel：用于开发半参数回归模型和定制贝叶斯推断算法的概率编程框架. arXiv: 2209.10975[stat]. 检索于2025年10月10日，网址：http://arxiv.org/abs/2209.10975 Santos, B., & Bolfarine, H. (2015). 使用分位数回归对零膨胀或一膨胀比例数据进行贝叶斯分析. Journal of Statistical Computation and Simulation, 85(17), 3579–3593. https://doi.org/10.1080/ 00949655.2014.986733 Stasinopoulos, M. D., Kneib, T., Klein, N., Mayr, A., & Heller, G. Z. (2024). 位置、尺度和形状广义加性模型：一种分布回归方法及其应用 (第56卷). 剑桥大学出版社. https://doi.org/10.1017/9781009410076 Waldmann, E., Kneib, T., Yue, Y. R., Lang, S., & Flexeder, C. (2013). 贝叶斯半参数加性分位数回归. Statistical Modelling, 13(3), 223–252. https://doi.org/10.1177/1471082X13480650 Wood, S. N. (2017). 广义加性模型：R语言入门 (第2版). Chapman; Hall/CRC. https: //doi.org/10.1201/9781315370279 Wood, S. N. (2025). Mgcv：带自动平滑度估计的混合广义加性模型计算工具 [R软件包版本1.9-3]. https://CRAN.R-project.org/package= mgcv Yue, Y., & Rue, H. (2011). 加性混合分位数回归模型的贝叶斯推断. Computational Statistics and Data Analysis, 55(1), 84–96. https://doi.org/10.1016/j.csda.2010.05.006