

基于非参数分位数估计的众数回归模型 *

刘婷婷 杨联强* 王学军

(安徽大学数学科学学院, 合肥, 230601)

摘要: 本文给出基于非参数分位数估计的众数回归模型. 不同于传统的均值和中位数回归模型, 众数回归模型使用条件众数刻画分布的中心, 在数据分布存在异常值、非对称或重尾这些特征时, 具有更好的稳健性. 当前众数回归模型的解法主要是基于条件密度的核估计方法. 本文给出一种新的基于非参数分位数估计的众数回归模型求解方法. 该方法通过分布函数与分位数函数的互逆性来估计众数, 从而可以充分利用非参数分位数估计的灵活性. 模拟和实际应用结果显示, 该方法表现良好, 优于基于线性分位数估计的众数回归模型.

关键词: 众数; 分位数; 核; 回归

中图分类号: O212.7

英文引用格式: LIU T T, YANG L Q, WANG X J. Modal regression based on nonparametric quantile estimator [J]. Chinese J Appl Probab Statist, 2020, 36(5): 483–492. (in Chinese)

§1. 引言

传统的均值和中位数回归模型通过条件均值和中位数来刻画条件分布的中心. 当该条件分布是单峰对称分布时, 两种模型表现良好. 但是, 当条件分布有偏、重尾或者多峰时, 均值和中位数并不能很好的代表该分布的中心, 此时众数是更合适的选择. 众数回归是用“最有可能出现”的值(众数)来刻画条件分布的“中心”, 因此对于同样区间长度, 条件众数比均值和中位数的预测区间包含更多的样本点; 另外, 众数回归突出的一个优点是比均值回归模型具有更好的稳健性. 真实数据往往存在多峰、非对称、重尾等现象, 离群值也时常存在, 正是这些特征, 使得众数回归模型被越来越多的使用在实际问题研究上, 例如, 健康食谱数据的分析^[1], 温度数据分析和预测^[2], 阿尔茨海默症分析和预测^[3]等工作.

关于单值众数(整体众数)函数的回归模型理论, 最早工作由 Sager 和 Thisted^[4] 在研究单调回归时提出, 并使用密度估计求解众数估计量; Lee^[5] 提出了线性众数回归模型, 该模型使用 0-1 损失函数, 求解方法为最大似然估计法. 然后, 其他学者对该方法作出了一系列改进工作, 例如 Yao 和 Li^[6] 提出线性众数回归模型, 并讨论了估计量的渐近正态性;

*国家自然科学基金项目(批准号: 11671012)、安徽省高校自然科学基金项目(批准号: KJ2017A028、KJ2017A024)和安徽大学数学科学学院开放课题(批准号: Y01002431)资助.

*通讯作者, E-mail: yanglq@ahu.edu.cn.

本文 2019 年 5 月 13 日收到, 2019 年 10 月 14 日收到修改稿.

Kemp 和 Silva^[7] 研究了半参数模型; Krief^[8] 提出了半线性众数回归模型等; 另外, Zhao 等^[9] 给出了线性众数函数的回归系数的经验似然估计; Lee 和 Kim^[10] 讨论了删失数据下的线性众数回归模型等. 而关于多值众数(局部众数)函数回归模型的理论研究需要更复杂的分析和计算方法, 研究成果尚不多见. 其中 Einbeck 和 Tutz^[11] 给出了基于核密度估计的众数回归模型系统理论, 并讨论了运用均值漂移算法求解估计量; Chen 等^[12] 基于均值漂移算法系统讨论了非参数众数回归估计量的渐近收敛性, 并构建了置信区间和预测区间; Chen 等^[13] 则讨论了基于该方法的聚类问题; Zhou 和 Huang^[14] 进一步将上文方法应用于存在测量误差问题的数据分析.

综合现有研究成果可见, 参数众数模型求解和理论性质分析已有较好成果, 但参数模型的局限在于函数表达能力受限. 非参数众数模型更富有变化性, 但解法和理论性质推导要复杂, 比较流行的解法有广义 EM 算法和均值漂移(mean shift)算法(参见文献 [15]). 但这两种方法的起点本质上都是核密度估计, 因此估计效果依赖于核密度估计的质量. 在不依赖于核密度估计的方法中, Ota 等^[16] 提出了一种基于线性分位数函数估计的众数回归模型, 将众数回归函数的估计问题转化为分位数函数导数最值的求解问题. 该方法将条件分位数函数设为线性函数, 所以是一种参数模型的方法. 本文在此基础上, 给出一种非参数的方法, 该方法充分利用非参数分位数估计优秀的函数的表达能力, 来提升对复杂众数函数的估计效果. 全文后续内容结构如下, 在第二节中介绍该模型的构造和求解, 第三节和第四节分别给出三个模拟研究和一个真实数据研究, 并与线性模型进行比较, 第五节对该方法的优点和不足作出总结.

§2. 基于分位数函数的众数回归

设 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ 是来自总体 (X, Y) 的简单随机样本, 记 $f(y | X = x)$ 为给定 $X = x$ 时 Y 的条件密度函数, 则众数回归函数定义为

$$m(x) = \text{Mode}(Y | X = x) = \arg \max_y f(y | X = x).$$

基于核密度估计的众数回归模型求解方法的研究比较充分, 详情可见综述性文献 [15]. 在文献 [16] 中, 作者使用了一种新的基于线性分位数函数的众数估计方法. 我们首先给出这种方法理论基础的证明, 然后给出基于非参数分位数估计的众数函数估计, 讨论模型超参数的选择方法并与线性模型比较.

2.1 基于分位数函数的众数函数估计

定理 1 记 $f(y | x)$ 是给定 $X = x$ 时 Y 的条件密度函数, $F(y | x)$ 为给定 $X = x$ 时 Y 的条件分布函数, $y = Q(\tau | x)$ 表示 $X = x$ 时 Y 的 τ 分位数函数, $D = \{y | f(y | x) > 0\}$.

设 $m(x)$ 为给定 $X = x$ 时 Y 的单值众数函数 (即仅考虑存在唯一的全局众数), 如果在 D 上, $\partial F(y|x)/\partial y = f(y|x)$ 且 $\tau_x = \arg \min_{\tau} \partial Q(\tau|x)/\partial \tau$, 那么 $m(x) = Q(\tau_x|x)$.

证明: 因为 $y = Q(\tau|x)$ 且 $F(y|x) = \tau$, 即条件分位数函数与条件分布函数互为反函数. 由众数函数定义及上述条件可知, 在区域 D 上,

$$\begin{aligned} m(x) &= \arg \max_y f(y|x) = \arg \max_y \frac{\partial F(y|x)}{\partial y} \\ &= \arg \max_y \frac{1}{\partial Q(\tau|x)/\partial \tau} = \arg \min_{Q(\tau|x)} \frac{\partial Q(\tau|x)}{\partial \tau}. \end{aligned}$$

又由 $\tau_x = \arg \min_{\tau} \partial Q(\tau|x)/\partial \tau$, 即得 $m(x) = Q(\tau_x|x)$. \square

由此定理知, 众数回归函数的求解问题可转化为分位数函数关于分位点导数的最小值求解问题, 这给我们提供了一种全新的不同于基于核密度估计的众数回归模型解法, 而且, 分位数函数估计已有丰富的研究成果可以借用. Ota 等^[16] 的方法是将分位数函数设为线性函数, 进而依照上述理论估计众数函数. 这种参数线性模型的优点是简洁直观, 可解释性强, 求解方便, 理论性质推导也较简洁, 但缺点是线性模型的函数表达能力有限, 在分位数函数变化复杂, 特征多样时, 容易出现模型设定错误. 因此, 本文运用基于核方法的非参数分位数函数估计方法, 进而给出众数回归模型的解的估计.

基于核方法的非参数分位数函数模型将模型的解设定在一个合适的再生核希尔伯特空间 H 中, 该空间足够大, 能保证解的良好逼近性质, 同时通过设定正则化项, 使得回归函数足够光滑. 然后, 通过著名的表示定理, 将模型解表示成核对应的特征函数的线性组合形式, 将模型求解问题转化为一个凸优化问题, 然后转化为对偶问题求解, 该方法的基础理论可参见文献 [17] 以及文献 [18], 而 Takeuchi 等^[19] 对该分位数估计方法进行了详细的介绍. 下面我们首先介绍线性分位数估计方法的实现, 然后介绍非参数分位数估计的实现过程.

分位数函数估计的线性模型方法是将分位数函数设为 $Q(\tau|X=x) = x^T \beta(\tau)$, $\tau \in (0, 1)$, 斜率 $\beta(\tau)$ 可以由下式估计:

$$\hat{\beta}(\tau) = \arg \min_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \rho_\tau(Y_i - X_i^T \beta),$$

其中 $\rho_\tau(\mu)$ 是检测函数, 也称作损失函数. 这是一个凸线性规划问题, 求解原理这里不再赘述, 可以通过 R 软件中的程序包 quantreg 直接实现.

下面介绍基于核方法的非参数分位数估计. 其中, 核方法及再生核希尔伯特空间可参考文献 [17] 的第 2 和第 4 章以及文献 [18] 的第 4 章, 基于核方法的回归函数估计可见文献 [17] 的第 9 章, 分位数函数估计可见文献 [18] 的第 9 章和文献 [19]. 方法简介如下, 记分位数回归的损失函数为

$$l_\tau(\xi) = \begin{cases} \tau \xi, & \xi \geq 0; \\ (\tau - 1)\xi, & \xi \leq 0. \end{cases}$$

核 $k(\cdot, \cdot)$ 对应的再生核希尔伯特空间为 H , 则分位数函数估计量为如下带正则化项优化问题的解:

$$\hat{f} = \arg \min_f \left\{ \frac{1}{n} \sum_{i=1}^n l_\tau(Y_i - f(X_i)) + \frac{\lambda}{2} \|g\|_H^2 \right\}, \quad (1)$$

其中 $f = g + b$, $b \in \mathbb{R}$, $g \in H$. $\|\cdot\|_H$ 是再生核希尔伯特空间 H 的范数. 根据表示定理(参见文献 [17] 的第 4.2 节以及文献 [18] 的第 5.1 和 5.2 节), 该优化问题的解一定满足

$$\hat{f}(x) = \sum_{j=1}^n \omega_j k(x, X_j) + b,$$

进一步得到该问题的等价形式:

$$\min_{w, b, \xi_i, \xi_i^*} C \sum_{i=1}^n \tau \xi_i + (1 - \tau) \xi_i^* + \mathbf{w}^\top \mathbf{K} \mathbf{w},$$

使得

$$\begin{aligned} y_i - \sum_{j=1}^n w_j k(X_i, X_j) - b &\leq \xi_i, \\ \sum_{j=1}^n w_j k(X_i, X_j) + b - y_i &\leq \xi_i^*, \end{aligned} \quad \xi_i, \xi_i^* \geq 0.$$

其中 $\mathbf{w} = (w_1, w_2, \dots, w_n)^\top$, $\mathbf{K} = (k(X_i, X_j))_{n \times n}$ 是 Gram 矩阵. 常用的核函数有高斯核 $k(u, v) = e^{-\|u-v\|^2/(2\sigma^2)}$, 拉普拉斯核 $k(u, v) = e^{-\|u-v\|/\sigma}$, 线性核 $k(u, v) = u^\top v + c$ 等. 然后, 上问题即可转化为对偶问题, 并使用二次规划方法求解. 整个求解过程可以通过 R 软件的程序包 kqr 实现(参见文献 [19]).

两种方法估计出来的分位数函数 $\hat{Q}_x(\tau) = \hat{f}(x)$ 关于 τ 都没有初等函数表示, 无法直接求其导数. 因此, 进一步使用差商来估计分位数 $Q_x(\tau)$ 关于 τ 的导数 $s_x(\tau)$, 即:

$$\hat{s}_x(\tau) = \frac{\hat{Q}_x(\tau + h) - \hat{Q}_x(\tau - h)}{2h}.$$

并令 $\hat{\tau}_x = \arg \min_\tau \hat{s}_x(\tau)$, 最后得到条件众数的估计 $\hat{m}(x) = \hat{Q}_x(\hat{\tau}_x)$.

2.2 超参数选择

基于线性分位数函数的众数模型需要选择计算差商的步长 h , 基于核方法分位数估计的众数回归模型需要确定两个超参数, 除了差商的步长 h 外, 还要选择正则化项权重参数 C . 两种方法中, 步长 h 的选择采用与 Ota 等^[16] 相同, 过程如下. 首先, 基于 Koenker 和 Machado^[20] 的方法计算

$$h^{\text{KM}}(\tau) = n^{-1/3} z_\alpha^{2/3} \left[\frac{3\phi(\Phi^{-1}(\tau))}{4\Phi^{-1}(\tau)^2 + 1} \right]^{1/3},$$

其中 ϕ 和 Φ 是标准正态分布的密度函数和分布函数, $z_\alpha = \Phi^{-1}(1 - \alpha/2)$, 设 $\alpha = 0.05$; 其次, 在给定的 $X = x$ 处, 使用一个试点带宽 $h^{\text{pilot}} = n^{1/6}h^{\text{KM}}(0.5)$ 计算一个初始估计 $\hat{\pi}_x^{\text{prelim}}$; 最后使用 $h_n = n^{1/6}h^{\text{KM}}(\hat{\pi}_x^{\text{prelim}})$ 估计 $m(x)$.

正则化项权重参数 C 的选择控制着分位数函数的光滑度, 从而对众数回归函数估计的影响较大. 在经典的均值回归模型中, 该参数选择方法一般是基于似然信息准则或交叉验证得分. Zhou 和 Huang^[21]的模拟计算研究表明, 核密度估计带宽的选择方法并不适合众数回归模型. 同样的, 通常的均值回归模型的带宽权重参数选择法并不适合众数回归. 因此, 本文根据众数回归的本质特征, 基于交叉验证思想, 构造一个新的交叉验证准则如下:

$$\text{CV}(C) = \frac{1}{n} \sum_{i=1}^n I\left(\frac{|Y_i - \hat{m}^{(-i)}(X_i)|}{d} \leq 1\right),$$

其中 $I(\cdot)$ 为示性函数, d 是常数, 取值为响应变量观测值 (Y_1, Y_2, \dots, Y_n) 极差的 5%, $\hat{m}^{(-i)}(\cdot)$ 为除去第 i 个观测值时回归函数的估计值, 搜索选出使 CV 取得最大值的参数 C . 模拟和真实数据分析显示该准则表现良好, 能选出合适的参数 C 的值.

§3. 模拟

本节通过三个模拟来演示基于非参数分位数估计的众数回归模型的估计效果, 并与 Ota 等^[16]给出的线性模型进行比较, 比较准则为真实众数函数与估计的众数函数在 X_i 处的均方误差

$$\text{MSE} = n^{-1} \sum_{i=1}^n [\hat{m}(X_i) - m(X_i)]^2.$$

在每个例子中, 差商估计的步长和正则化权重参数的选择均按 2.2 节方法进行. 其中, 正则化参数 C 对应的交叉验证得分 $\text{CV}(C)$ 取值如图 1 所示, 例 2 最优的 C 值为 2.2 (图 1(a)), 例 3 最优的 C 值为 0.7 (图 1(b)), 例 4 最优的 C 值为 0.1 (图 1(c)). 估计结果分别见图 2、3、4.

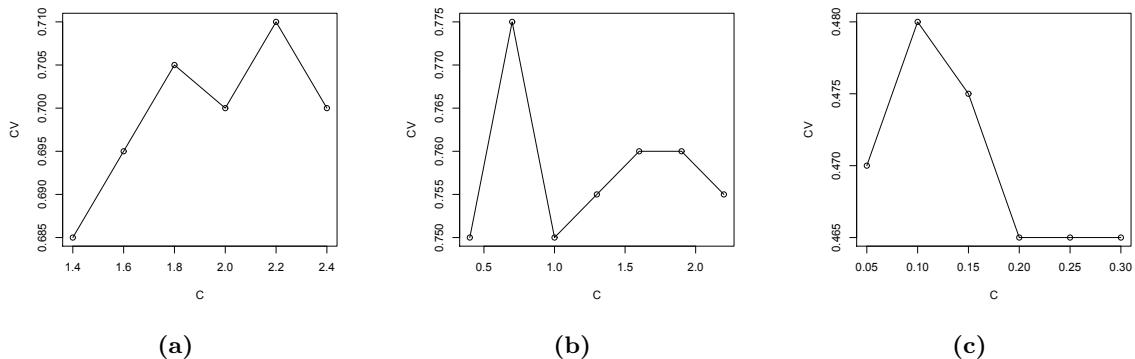


图 1 正则化参数 C 与交叉验证得分 $\text{CV}(C)$ 变化图. (a) 例 2; (b) 例 3; (c) 例 4

例 2 令

$$f(x) = 50x^3 - 75x^2 + \frac{75}{2}x - \frac{25}{4},$$

$$Y_i = f(X_i) + \varepsilon_i, \quad X_i \sim U(0, 1),$$

$$\varepsilon_i \sim 0.2N(-2, 0.5^2) + 0.8N(2, 0.5^2),$$

样本容量 $n = 200$. 设 $x_1 = x^3$, $x_2 = x^2$, $x_3 = x$, F 表示 ϵ 的分布函数, 则真实的分位数函数记作线性函数为

$$Q_\tau(x) = 50x_1 - 75x_2 + \frac{75}{2}x_3 - \frac{25}{4} + F^{-1}(\tau),$$

ϵ 的众数为 2, 所以真实众数函数为

$$m(x) = 50x^3 - 75x^2 + \frac{75}{2}x - \frac{17}{4}.$$

估计的众数函数结果如图 2 所示, 其中线性模型的 MSE 为 0.03096441, 非参数模型的 MSE 为 0.06124879.

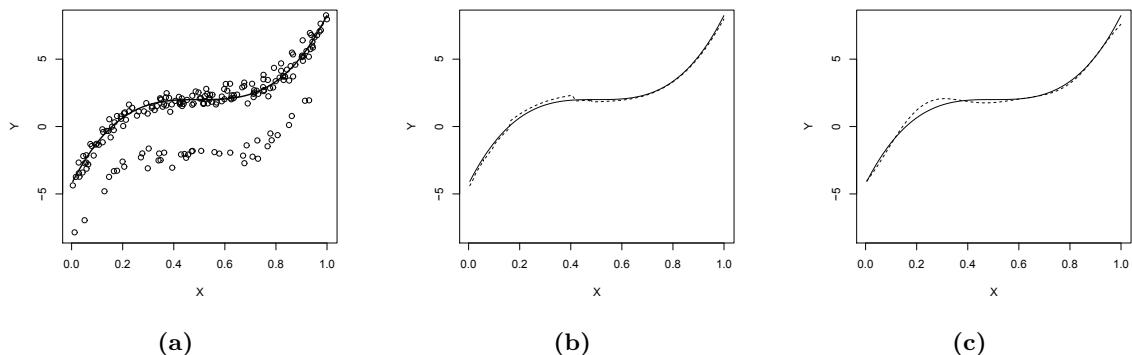


图 2 多项式众数函数、散点及其估计值. (a) 真实众数函数和散点图; (b) 真实众数函数 (实线) 及基于线性模型的回归函数 (虚线); (c) 真实众数函数 (实线) 及基于非参数模型的回归函数 (虚线)

例 3 令

$$f(x) = 8 \sin(10x), \quad Y_i = f(X_i) + \varepsilon_i, \quad X_i \sim U(0, 1),$$

$$\varepsilon_i \sim 0.2N(-4, 0.5^2) + 0.8N(4, 0.5^2),$$

样本容量 $n = 200$. 设 $x_1 = \sin(10x)$, F 表示 ϵ 的分布函数, 则真实的分位数函数的线性形式为

$$Q_\tau(x) = 8x_1 + F^{-1}(\tau),$$

ϵ 的众数为 4, 所以真实众数函数为

$$m(x) = 8 \sin(10x) + 4.$$

估计的众数函数结果如图 3 所示, 其中线性模型的 MSE 为 0.3814918, 非参数模型的 MSE 为 0.1420147.

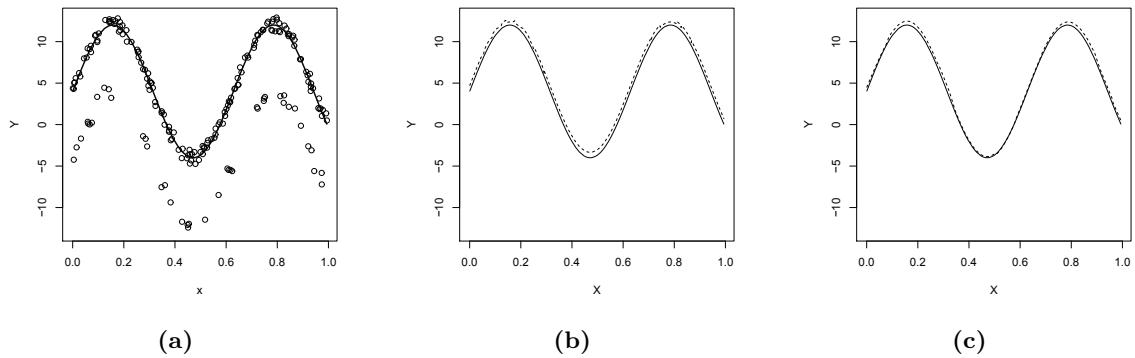


图 3 周期众数函数、散点及其估计图像. (a) 真实众数函数和散点图; (b) 真实众数函数 (实线) 及基于线性模型的回归函数 (虚线); (c) 真实函数 (实线) 及基于非参数模型的回归函数 (虚线)

例 4 (Mexican Hat) 令

$$f(x) = -1 + 1.5x + 0.2\phi(x - 0.6), \quad Y_i = f(X_i) + \varepsilon_i,$$

ϕ 表示均值为 0, 标准差为 0.04 的密度函数,

$$X_i \sim U(0, 1), \quad \varepsilon_i \sim 0.4N(-1, 2.5^2) + 0.6N(1, 0.1^2),$$

样本容量 $n = 200$. 设 $x_1 = x$, $x_2 = \phi(x - 0.6)$, F 表示 ϵ 的分布函数, 则真实分位数函数的线性形式为

$$Q_\tau(x) = -1 + 1.5x_1 + 0.2x_2 + F^{-1}(\tau),$$

ϵ 的众数为 1, 所以真实众数函数为

$$m(x) = 1.5x_1 + 0.2x_2.$$

估计的众数函数结果如图 4 所示, 其中线性模型的 MSE 为 0.01068117, 非参数模型的 MSE 为 0.02123267.

由以上三个模拟可以看出, 基于非参数分位数估计的众数回归模型能良好的估计出众数回归函数, 估计效果与基于正确设定线性形式的线性分位数模型的估计效果接近. 但必

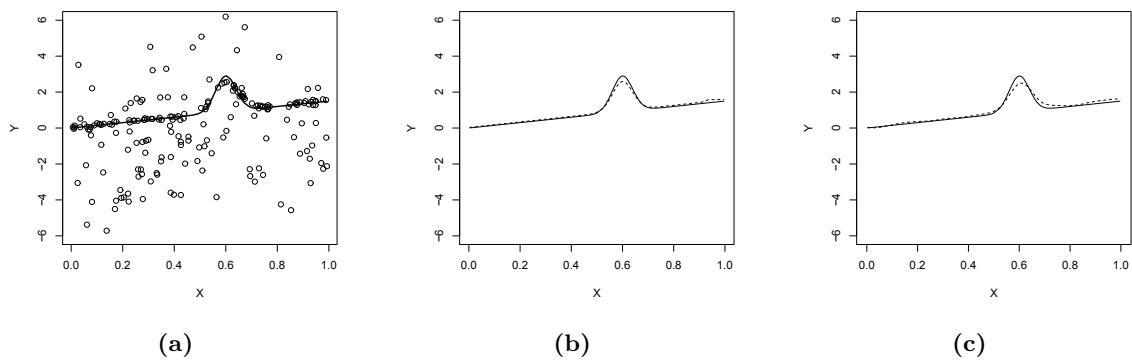


图 4 Mexican Hat 众数函数、散点及其估计图像. (a) 真实众数函数和散点图; (b) 真实众数函数 (实线) 及基于线性模型的回归函数 (虚线); (c) 真实函数 (实线) 及基于非参数模型的回归函数 (虚线)

须指出的是, 这是因为在模拟中, 我们知道真实分位数函数的具体形式, 从而可以将其设置成对应的完全准确的线性形式, 这相当于对线性模型估计的准确性提供了极大而且准确的先验信息, 因此基于线性模型方法的表现也非常好. 但是, 在实际应用中, 不可能知道分位数函数正确的线性形式, 这样会使得基于线性分位数模型的估计效果显著变差, 这一点在下一节的实际应用中体现得非常明显. 而基于非参数分位数估计的众数模型却不受此影响, 仍然能良好的估计众数函数.

§4. 应用

本节将前文介绍的两种众数模型运用于老忠实喷泉 (Old Faithful) 的数据. 该数据包含 272 个老忠实泉的喷发持续时间和随后的等待时间, 数据取自 R 程序包 faithful. 我们以等待时间为自变量, 以喷发时间为响应变量. 估计结果见图 5. 由图 5 可以看出, 基于线性分位数估计的众数回归函数拟合结果过于简单, 不能很好的描述两个变量间的相依关系, 而且拟合的回归函数有难以消除的跳跃, 其根本原因就是线性模型的设定错误. 但基于非参数分位数函数的众数回归模型估计效果良好, 回归曲线较光滑, 且能较好的刻画两变量间的相依变化特征.

§5. 总结

本文提出了一种基于非参数分位数估计的众数回归模型, 通过使用基于核方法的分位数函数估计, 再利用分位数函数与众数的关系求出众数回归函数的估计. 我们还构造了一种适合众数回归模型超参数选择的交叉验证准则, 并据此给出了模拟和应用实例. 结果显示该非参数众数回归模型表现良好, 相比于基于线性分位数函数的众数回归模型, 该方法

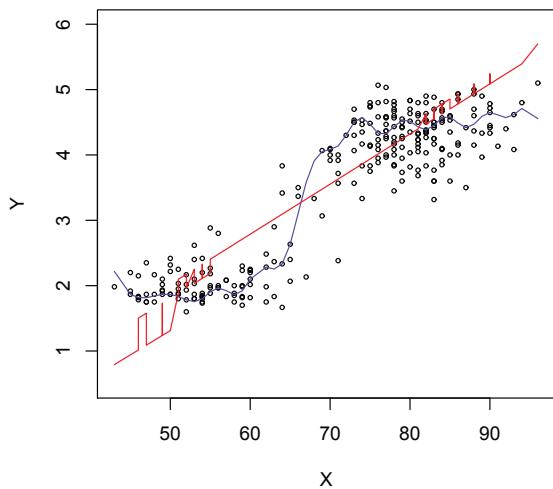


图5 老忠实泉数据散点, 基于线性分位数估计的众数回归函数(红实线)
和基于非参数分位数方法的众数回归函数(蓝实线)

更具灵活性, 能更好的匹配具有复杂变化特征的数据集. 但是, 该估计结果的渐近性质分析, 以及基于此估计结果的对总体众数的统计推断问题, 是需要进一步研究的工作.

参 考 文 献

- [1] ZHOU H M, HUANG X Z. Nonparametric modal regression in the presence of measurement error [J]. *Electron J Statist*, 2016, **10**(2): 3579–3620.
- [2] HYNDMAN R J, BASHTANNYK D M, GRUNWALD G K. Estimating and visualizing conditional densities[J]. *J Comput Graph Statist*, 1996, **5**(4): 315–336.
- [3] WANG X Q, CHEN H, CAI W D, et al. Regularized modal regression with applications in cognitive impairment prediction [J]. *Adv Neural Inform Process Syst*, 2017, **30**: 1448–1458.
- [4] SAGER T W, THISTED R A. Maximum likelihood estimation of isotonic modal regression [J]. *Ann Statist*, 1982, **10**(3): 690–707.
- [5] LEE M J. Mode regression [J]. *J Econometrics*, 1989, **42**(3): 337–349.
- [6] YAO W X, LI L H. A new regression model: modal linear regression [J]. *Scand J Stat*, 2014, **41**(3): 656–671.
- [7] KEMP G C R, SILVA J M C S. Regression towards the mode [J]. *J Econometrics*, 2012, **170**(1): 92–101.
- [8] KRIEF J M. Semi-linear mode regression [J]. *Econom J*, 2017, **20**(2): 149–167.
- [9] ZHAO W H, ZHANG R Q, LIU Y K, et al. Empirical likelihood based modal regression [J]. *Statist Papers*, 2015, **56**(2): 411–430.
- [10] LEE M J, KIM H. Semiparametric econometric estimators for a truncated regression model: a review with an extension [J]. *Stat Neerl*, 1998, **52**(2): 200–225.

- [11] EINBECK J, TUTZ G. Modelling beyond regression functions: an application of multimodal regression to speed-flow data [J]. *J R Stat Soc Ser C Appl Stat*, 2006, **55**(4): 461–475.
- [12] CHEN Y C, GENOVESE C R, TIBSHIRANI R J, et al. Nonparametric modal regression [J]. *Ann Statist*, 2016, **44**(2): 489–514.
- [13] CHEN Y C, GENOVESE C R, WASSERMAN L. A comprehensive approach to mode clustering [J]. *Electron J Statist*, 2016, **10**(1): 210–241.
- [14] ZHOU H M, HUANG X Z. Nonparametric modal regression in the presence of measurement error [J]. *Electron J Statist*, 2016, **10**(2): 3579–3620.
- [15] CHEN, Y C. Modal regression using kernel density estimation: a review [J]. *Wiley Interdiscip Rev Comput Stat*, 2018, **10**(4): e1431, 14 pages.
- [16] OTA H, KATO K, HARA S. Quantile regression approach to conditional mode estimation [J]. *Electron J Statist*, 2019, **13**(2): 3120–3160.
- [17] SCHÖLKOPF B, SMOLA A J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* [M]. Cambridge, MA: MIT Press, 2001.
- [18] STEINWART I, CHRISTMANN A. *Support Vector Machines* [M]. New York: Springer-Verlag, 2008.
- [19] TAKEUCHI I, LE Q V, SEARS T D, et al. Nonparametric quantile estimation [J]. *J Mach Learn Res*, 2006, **7**: 1231–1264.
- [20] KOENKER R, MACHADO J A F. Goodness of fit and related inference processes for quantile regression [J]. *J Amer Statist Assoc*, 1999, **94**(448): 1296–1310.
- [21] ZHOU H M, HUANG X Z. Bandwidth selection for nonparametric modal regression [J]. *Comm Statist Simulation Comput*, 2019: **48**(4): 968–984.

Modal Regression Based on Nonparametric Quantile Estimator

LIU Tingting YANG Lianqiang WANG Xuejun

(School of Mathematical Science, Anhui University, Hefei, 230601, China)

Abstract: Modal regression based on nonparametric quantile estimator is given. Unlike the traditional mean and median regression, modal regression uses mode but not mean or median to represent the center of a conditional distribution, which helps the model to be more robust for outliers, asymmetric or heavy-tailed distribution. Most of solutions for modal regression are based on kernel estimation of density. This paper studies a new solution for modal regression by means of nonparametric quantile estimator. This method builds on the fact that the distribution function is the inverse of the quantile function, then the flexibility of nonparametric quantile estimator is utilized to improve the estimation of modal function. The simulations and application show that the new model outperforms the modal regression model via linear quantile function estimation.

Keywords: mode; quantile; kernel; regression

2010 Mathematics Subject Classification: 62G08