【统计理论与方法】

基于贝叶斯方法的比例数据分位数推断及其应用

赵为华[1]，张日权[2]
(1. 南通大学 理学院，江苏 南通 226019；2. 华东师范大学 统计学院，上海 200241)

摘要：为了尝试使用贝叶斯方法研究比例数据的分位数回归统计推断问题，首先基于 Tobit 模型给出了分位数回归建模方法，然后通过选取合适的先验分布得到了贝叶斯层次模型，进而给出了各参数的后验分布并用于 Gibbs 抽样。数值模拟分析验证了所提出的贝叶斯推断方法对于比例数据分析的有效性。最后，将贝叶斯方法应用于美国加州海洛因吸毒数据，在不同的分位数水平下揭示了吸毒频率的影响因素。

关键词：比例数据；分位数回归；Tobit 模型；贝叶斯分析；Gibbs 抽样
中图分类号：C812；O212 **文献标志码**：A **文章编号**：1007－3116(2016)08－0009－05

一、引 言

比例数据在很多实际问题中大量存在，如家庭食物消费支出占家庭收入的比例，某种疾病的临床治愈率，公司某项产品的市场占有率，银行到期贷款的按期偿还比例等。对比例数据的回归建模方法已有一些研究，如 Peter 和 Tan 基于单纯形分布研究了连续比例数据的回归参数估计问题[1]；Ferrari 等通过参数变换的方法，利用 Beta 回归模型并借鉴广义线性模型的理论和方法研究了比例数据的估计及其诊断分析问题[2]；Papke 和 Wooldridge 提出了基于拟似然方法研究比例数据的回归建模问题[3]；其他的一些相关工作详见 Kieschnick、Zhao、Ramalho 等人的研究成果[4-8]。

上述研究都是基于均值回归分析方法展开的，然而均值回归只能提供在平均意义下比例因变量与自变量之间的相关关系且易受异方差或异常值影响。在实例分析中，应用者希望全方位、多角度分析和揭示因变量与自变量之间的关系，分位数回归就能提供关于因变量的条件分布在不同分位点下的全面描述，进而能够得到在低、中、高等分位数下自变量影响因变量的动态变化关系。分位数回归最早是由 Koenker 等于 1978 年在研究计量经济建模时提出的统计方法，与基于均值回归相比，分位数回归不受数据中异常点的影响，是一种稳健的估计方法，特别当分位数 $\tau=0.5$ 时即为中位数回归(俗称最小一乘方法)[9]。此外，分位数回归不需要 Gauss-Markov 条件假定，既能适应误差方差无穷的情形，又能方便地处理异方差、多峰数据等情形。关于分位数回归的详细内容可参见 Koenker 等人的专著[10]。

目前关于比例数据的分位数回归建模及其统计推断研究刚刚开始，主要原因是其估计的相关理论性质难以建立，以及如何将得到的大样本理论性质应用到实际数据分析中。相比于经典的频率方法，贝叶斯方法具有如下优点：贝叶斯方法原理简单且易实施；通过选取合适的参数先验，基于后验分布除了可以得到 Bayesian 后验估计值，同时可以得到 Bayesian 可信区间，这为我们提供一种选择相关重要自变量的方法；很多时候，特别在小样本时，由贝叶斯方法得到的结果要比基于频率方法得到的结果更可靠。为此，本文拟应用贝叶斯方法研究比例数

收稿日期：2016－02－28
基金项目：教育部人文社科青年基金项目《比例数据的分位数回归建模》(14YJC910007)；国家自然科学基金项目《函数型含指标项半参数回归模型的统计分析》(11571112)
作者简介：赵为华，男，江苏海门人，统计学博士，副教授，硕士生导师，研究方向：回归建模及其应用；
张日权，男，山西大同人，统计学博士，教授，博士生导师，研究方向：分位数回归和半参数模型。

据的分位数推断问题。

二、比例数据的分位数回归建模

由于比例数据的有界性即取值在[0,1]上，因此直接基于 Koenker 的分位数回归建模方法会失效，得到的预测估计往往会超出上、下界。下面基于 Tobit 模型引进比例数据的分位数回归建模方法[10-12]。Tobit 模型最早用来刻画因变量取值有上限或者下限时或者有极限值时提出来的，已在许多领域特别是计量经济学领域中有着广泛应用。经典的 Tobit 线性回归模型假定因变量取值有下界，并通过引入一个不可观测的潜在变量后模型可表示为：

$$\begin{aligned} y^* &= x^T\beta+\varepsilon \\ y &= \max(y_0, y^*) \end{aligned} \tag{1}$$

其中 y 是一个观测因变量，y_0 是观测因变量的取值下限，y^* 是潜在因变量，ε 是误差项，$x=(x_1,x_2,\cdots,x_p)^T\in R^p$ 是 p 维自变量，$\beta=(\beta_1,\beta_2,\cdots,\beta_p)^T$ 是回归系数。如果假定协变量中的第一个分量 $x_1\equiv 1$，则模型(1)中包含了常数项。Tobit 回归本质上是一种具有固定删失点的模型，国内外许多文献在均值回归和分位数回归下分别对 Tobit 线性回归模型展开过系统研究，建立了相关的大样本理论性质，并将研究结果应用于相关计量经济学问题。近年来，由于统计软件的计算能力快速提高以及贝叶斯方法的优点，一些文献基于贝叶斯方法研究了 Tobit 参数回归模型[13]，Kaifeng Zhao 等进一步研究了 Tobit 半参数回归模型贝叶斯推断方法[14]。

由于比例数据既有下界 0 又有上界 1，基于模型(1)我们提出如下的 Tobit 模型：

$$\begin{aligned} y^* &= x^T\beta+\varepsilon \\ y &= y^*\cdot I(0<y^*<1)+I(y^*\geqslant 1) \end{aligned} \tag{2}$$

其中 $I(\cdot)$ 是示性函数。假设有 n 个独立观测 (x_i,y_i)，$i=1,2,\cdots,n$，借鉴已有关于 Tobit 分位数回归模型估计方法，系数 β 的估计可以通过极小化以下式子得到：

$$\hat{\beta}=\arg\min_{\beta}\sum_{i-1}^{n}\rho_\tau(y_i-\max(0,\min(1,x_i^T\beta))) \tag{3}$$

其中 $\rho_\tau(u)=u(\tau-I(u<0))$ 是分位数损失函数，$0<\tau<1$ 是分位数水平。尽管 Tobit 模型为比例数据的分位数回归建模提供了一条非常可行的途径，它可以处理连续型比例数据(因变量取值在开区间(0,1)时情形)，半连续型比例数据(因变量取值在闭区间[0,1]或半开半闭区间[0,1)或(0,1]时情形)，然而，直接基于式(3)不仅讨论系数估计的大样本理论性质是困难的，而且如何提出有效、快速的计算方法也是充满挑战的。由于贝叶斯统计的优点，下面我们基于贝叶斯方法提出比例数据的分位数层次模型，并基于层次模型提出各参数的 Gibbs 抽样方法。

三、贝叶斯层次模型和 Gibbs 抽样

贝叶斯推断需要似然函数，借鉴分位数回归贝叶斯推断方法，假设式(2)中的 y_i^* 在给定 $\mu_i=x_i^T\beta$ 时服从具有位置参数为 μ，尺度参数为 σ 的不对称拉普拉斯分布，其密度函数为：

$$\pi(y_i^*\mid\mu_i,\sigma)=\frac{\tau(1-\tau)}{\sigma}\exp\left\{-\frac{1}{\sigma}\rho_\tau(y_i^*-\mu_i)\right\} \tag{4}$$

其中 $0<\tau<1$ 称为偏度参数。进一步，式(4)可以等价地表示成如下的混合密度函数：

$$y_i^*=\mu_i+k_1e_i+\sqrt{k_2\sigma e_i}z_i \tag{5}$$

其中 $k_1=\dfrac{1-2\tau}{\tau(1-\tau)}$，$k_2=\dfrac{2}{\tau(1-\tau)}$，$e_i$ 服从参数为 $1/\sigma$ 的指数分布，z_i 服从标准正态分布。结合式(2)和式(5)，并对各个参数选取合适的先验分布，我们有如下的层次模型：

$$\begin{cases} y_i=y_i^*\cdot I(0<y_i^*<1)+I(y_i^*\geqslant 1) \\ y^*\mid\mu,e,\sigma\sim N(\mu+k_1e,E) \\ e_i\sim\exp(1/\sigma) \\ \beta\sim\prod_{j-1}^{p}\dfrac{\lambda}{2\sigma}\exp\left(-\dfrac{\lambda}{\sigma}\mid\beta_j\mid\right) \\ \sigma\sim IGa(a_1,b_1) \\ \lambda\sim Ga(a_2,b_2) \\ i=1,2,\cdots,n \end{cases} \tag{6}$$

其中 $y^*=(y_1^*,y_2^*,\cdots,y_n^*)^T$，$\mu=(\mu_1,\mu_2,\cdots,\mu_n)^T$，$e=(e_1,e_2,\cdots,e_n)^T$，$E=k_2\sigma\cdot\mathrm{diag}(e_1,e_2,\cdots,e_n)$，Ga 和 IGa 分别表示伽马和逆伽马分布，$a_1$、$b_1$、$a_2$、$b_2$ 是四个超参数，在模拟和实例中我们均设它们为 0.1。

注意到，在层次模型(6)中，我们取参数 β 的先验为拉普拉斯分布，尽管可以选取其他先验，如高斯先验，但选取拉普拉斯先验的好处在于上述的层次模型相当于 Bayesian Lasso 估计，具有缩减估计即变量选择的功能。为使得参数 β 的后验分布易于抽样，应用 Andrews 和 Mallows 恒等式[15]，$\dfrac{a}{2}e^{-a|t|}=\int_0^\infty\dfrac{1}{\sqrt{2\pi s}}e^{-\frac{t^2}{2s}}\dfrac{a^2}{2}e^{-\frac{a^2}{2}s}\mathrm{d}s$，$a>0$，则 β_j 的拉普拉斯先验

$\pi(\beta_j \mid \lambda,\sigma)=\frac{\lambda}{2\sigma}\exp\left(-\frac{\lambda}{\sigma}\mid\beta_j\mid\right)$亦可等价地表示为：

$$\beta_j \mid s_j,\lambda,\sigma \sim N(0,s_j),s_j \mid \lambda,\sigma \sim \exp((\lambda/\sigma)^2/2),\ j=1,2,\cdots,p \tag{7}$$

根据层次模型(6)以及式(7)，我们可以得到如下各参数的后验分布：

1. $y_i^* \mid \beta,y_i,\mu_i,\sigma,e_i \sim y_i I(0<y_i<1)+\mathrm{TN}_{(-\infty,0]}(\mu_i+k_1e_i,k_2\sigma e_i)\cdot I(y_i\leqslant 0)+\mathrm{TN}_{[1,+\infty)}(\mu_i+k_1e_i,k_2\sigma e_i)\cdot I(y_i\geqslant 1)$，其中 $\mathrm{TN}_{(-\infty,0]}$ 和 $\mathrm{TN}_{[1,+\infty)}$ 分别表示右截尾和左截尾正态分布。

2. $\beta \mid y^*,\mu,\lambda,\sigma,e,s \sim N(\eta,\Sigma)$，其中 $\Sigma=(G+X^THX)^{-1},\eta=\Sigma HX^Ty^*,G=\mathrm{diag}(1/s_1,1/s_2,\cdots,1/s_n),s=(s_1,s_2,\cdots,s_p)^T,H=\mathrm{diag}(1/k_2\sigma e_1,1/k_2\sigma e_2,\cdots,1/k_2\sigma e_n),X=(x_1,x_2,\cdots,x_n)^T,x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T$。

3. $s_j \mid \beta_j,\lambda,\sigma \sim \mathrm{GIG}(0.5,\beta_j^2,\lambda^2/\sigma^2)$，其中 GIG 表示广义逆高斯分布。

4. $e_i \mid \beta,\sigma,y^* \sim \mathrm{GIG}(0.5,(y_i^*-x_i^T\beta)^2/k_2\sigma,k_1^2+2/\sigma)$。

5. $\lambda \mid \beta,\sigma \sim \mathrm{Ga}(p+a_2,b_2+\sum_{j-1}^{p}\mid\beta_j\mid/\sigma)$。

6. $\sigma \mid \beta,\lambda \sim I\mathrm{Ga}(\frac{3}{2}n+p+a_1,b_1+\lambda\sum_{j-1}^{p}\mid\beta_j\mid)$。

根据各参数的后验分布，应用 Gibbs 抽样方法即可得到各参数的贝叶斯后验估计，如后验均值、后验中位数、后验标准差以及贝叶斯可信区间估计等。在 Gibbs 抽样中，我们每次模拟抽样15 000次，然后，为消除初值的影响剔除前面 5 000 次的抽样。下面的模拟分析和实例应用充分说明了本文给出的贝叶斯推断的有效性。

四、模拟研究

模拟数据由以下 Tobit 模型生成，$y^*=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+0.5\cdot\varepsilon,y=y^*\cdot I(0<y^*<1)+I(y^*\geqslant 1)$，其中 $\beta_0=0.5,\beta_1=1.2,\beta_2=-1,\beta_3=\beta_4=0;x=(x_1,x_2,x_3,x_4)^T$ 服从多元正态分布，x_j 的边际分布为标准正态且 $\mathrm{corr}(x_j,x_k)=0.5,1\leqslant j,k\leqslant 4$；随机误差 ε 服从标准正态分布 $N(0,1)$ 或者自由度为 4 的厚尾 T 分布 $t(4)$。根据模型的设置，模型中有两个重要变量 x_1 和 x_2，两个不相干变量 x_3 和 x_4。考虑样本容量为 $n=100,200$ 和 300，在不同的分位数水平 $\tau=0.1,0.25,0.5,0.75,0.9$ 下，每一种情况下均重复 500 次，回归系数后验平均估计和 95% 可信区间估计见表 1。

从表 1 可以看出，在正态误差下和厚尾 T 分布下，系数的估计值都非常接近于真值，而且对应的区间估计随着样本量的增加，区间的长度在不断地变小。同时，我们注意到对于两个不相干变量 x_3 和 x_4 的系数估计，其 95% 的可信区间始终包含 0。不难看出，本文给出的 Gibbs 抽样方法是令人满意的，而且对厚尾误差分布，估计的结果也是非常稳健的。此外，为了诊断 Gibbs 抽样的收敛性是否依赖于一些超参数的选择，在标准正态误差下，样本量 $n=300$ 时，超参数 $a_1=b_1=0.1$ 和 $a_1=b_1=0.5$ 情形下回归参数的 MCMC 轨迹图①。在轨迹图中我们发现其后验分布在不同初值下迅速地达到平稳分布，验证了本文所给的 Gibbs 抽样方法具有很好的收敛性且不依赖于初始参数的选择。

五、实例应用

本文将比例数据的贝叶斯分位数方法应用到一个社会学问题中。数据来自于美国加利福尼亚州的一个防范公民毒品瘾君子研究机构(Civil Addict Program)，该研究主要为吸食了海洛因毒品上瘾后有刑事犯罪记录的瘾君子进行强制药物治疗，然后评价强制药物治疗是否对瘾君子之后控制吸食海洛因的使用频率以及其他一些问题产生积极影响。该项目持续跟踪调查研究了 15 年，总共有 437 个有效样本。以 y 表示最后一年中每月平均使用海洛因 的频率(每月使用海洛因的天数/30 天)，解释变量主要包括跟踪的 15 年间吸毒人员每年进行药物治疗的平均月数(x_1)，15 年间每年被监禁的平均月数(x_2)，首次接受治疗时的年龄(x_3)和首次吸毒时的年龄(x_4)。因变量中大约有 40.5% 取值为 0，有24.9% 取值为 1，数据分布具有明显的不对称性和非常大的概率在区间端点取值，因此本文建议的比例数据回归模型非常适合此数据。考虑到有些解释变量取值的稀疏性并尽量消除解释变量取值的量纲影响，我们将 x_3 和 x_4 取值变换到单位区间上，将 x_1 和 x_2 变为取值离散化成 0 ～ 1 变量(按取值是否大于 6 进行划分)。在不同的分位数水平下，系数的估计及其 95% 可信区间估计拟合结果见表 2。

① 有需要了解的读者请与作者联系。

表 1　不同分位数水平下回归系数的点估计及其可信区间估计

样本量	误差	分位数 τ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$n=100$	$N(0,1)$	0.1	1.199 2	−1.012 3	0.011 6	0.003 9
			[0.903 1, 1.585 6]	[−1.356 1, −0.746 1]	[−0.152 5, 0.177 0]	[−0.170 8, 0.167 9]
		0.25	1.238 3	−1.039 6	0.019 2	0.002 9
			[0.930 2, 1.642 0]	[−1.394 2, −0.7610]	[−0.159 5, 0.1993]	[−0.170 0, 0.178 3]
		0.5	1.239 6	−1.087 3	0.005 3	0.012 7
			[0.963 4, 1.727 8]	[−1.471 7, −0.796 9]	[−0.177 9, 0.189 0]	[−0.166 0, 0.192 8]
	$t(4)$	0.1	1.181 2	−0.989 9	0.010 1	−0.002 1
			[0.831 9,1.633 4	[−1.393 7,−0.677 7]	[−0.213 5,0.236 3]	[−0.220 4,0.216 8]
		0.25	1.249 1	−1.041 1	0.009 6	0.007 7
			[0.902 5,1.709 6]	[−1.451 7,−0.729 0]	[−0.201 4,0.221 4]	[−0.204 3,0.220 0]
		0.5	1.261 4	−1.055 4	−0.007 7	0.012 3
			[0.912 7,1.726 3]	[−1.464 0,−0.744 2]	[−0.216 8,0.199 3]	[−0.195 8,0.221 5]
$n=200$	$N(0,1)$	0.1	1.184 7	−0.973 6	0.001 3	0.003 3
			[0.954 6, 1.411 7]	[−1.191 5,−0.788 3]	[−0.114 4, 0.111 8]	[−0.109 9, 0.120 6]
		0.25	1.209 6	−1.008 4	0.002 9	0.002 4
			[0.990 4, 1.479 2]	[−1.245 9, −0.811 8]	[−0.117 0, 0.123 9]	[−0.119 8, 0.123 4]
		0.5	1.204 2	−1.006 8	0.003 1	0.002 4
			[0.987 6, 1.467 6]	[−1.239 1, −0.813 7]	[−0.118 5, 0.125 5]	[−0.117 7,0.122 9]
	$t(4)$	0.1	1.167 4	−0.975 8	−0.007 8	0.007 4
			[0.958 4, 1.417 0]	[−1.194 9, −0.790 1	[−0.125 6, 0.109 4]	[−0.111 7, 0.127 5]
		0.25	1.206 3	−1.004 8	0.006 6	−0.006 9
			[0.992 6, 1.458 3]	[−1.226 1, −0.816 4]	[−0.110 1, 0.123 9]	[−0.125 5, 0.111 7]
		0.5	1.204 0	−1.005 3	0.002 4	−0.002 7
			[0.988 3, 1.461 8]	[−1.232 7, −0.814 2]	[−0.117 0, 0.121 7]	[−0.122 6, 0.117 6]
$n=300$	$N(0,1)$	0.1	1.154 4	−0.981 2	−0.001 0	0.003 1
			[0.984 7,1.348 9]	[−1.131 5,−0.811 7]	[−0.091 8,0.093 9]	[−0.096 1,0.090 2]
		0.25	1.205 1	−1.002 2	0.000 1	−0.002 0
			[1.025 9,1.411 4]	[−1.183 2,−0.844 6	[−0.096 5,0.096 3]	[−0.098 2,0.093 9]
		0.5	1.204 1	−1.001 1	−0.000 5	0.004 2
			[1.035 1,1.422 2]	[−1.194 0,−0.853 5]	[−0.097 1,0.096 7]	[−0.092 4,0.100 7]
	$t(4)$	0.1	1.161 9	−0.968 5	0.009 6	−0.014 5
			[0.958 8,1.398 4]	[−1.176 4,−0.787 4]	[−0.110 6,0.129 9]	[−0.136 4,1.07 5]
		0.25	1.215 2	−1.003 2	−0.005 3	−0.000 2
			[1.016 3,1.446 6]	[−1.205 7,−0.828 4]	[−0.119 8,0.109 7]	[−0.114 0,0.113 4]
		0.5	1.234 9	−1.021 6	−0.001 6	0.008 3
			[1.036 2,1.467 8]	[−1.227 6,−0.845 4]	[−0.113 3,0.110 6]	[−0.102 5,0.120 1]

表 2　海洛因吸毒数据的参数估计和 95%可信区间估计

分位数 τ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
0.25	0.140 6	−0.067 5	−0.292 9	−0.527 2	−0.441 1
	[−0.053 5,0.359 2]	[−0.261 5,0.098 1]	[−0.483 1,−0.120 9]	[−0.941 6,−0.161 3]	[−0.999 2,−0.021 4]
0.5	0.582 3	−0.173 0	−0.268 0	−0.488 7	−0.640 8
	[0.294 5, 0.858 3]	[−0.381 9, 0.029 3]	[−0.465 8,−0.074 3]	[−0.962 3,−0.034 3]	[−1.363 8,−0.013 0]
0.75	1.489 0	−0.355 7	0.139 0	−0.777 5	−1.145 9
	[1.168 1,1.826 9]	[−0.654 6,−0.041 2]	[−0.107 0,0.387 6]	[−1.492 5,−0.051 3]	[−1.852 8,−0.467 2]

从表 2 中我们有以下几个发现：第一，在三个不同的低、中、高分位数水平下，$\hat{\beta}_1$ 的估计值是负的，说明对瘾君子进行药物治疗每年大于 6 个月要比小于 6 个月的吸食海洛因的频率有下降趋势，特别对于高频率吸毒者($\tau=0.75$)具有明显的效果，其系数的可信区间不包含 0；第二，在中、低分位数水平下，吸毒者每年被监禁的平均月数对控制其使用海洛因的频率具有积极效果，而对于高频率毒品吸食者没有显著效果；第三，对于首次接受治疗时的年龄和首次吸毒时的年龄而言，它们的系数估计在三个分位数水平下都是负的，且各自的 95%可信区间均不包含 0，说明这两个因素是影响吸毒频率的重要因素。我们发现首次吸毒时的年龄越低，其以后使用海洛因的频率具有越高的倾向，因此要对青少年提早进行干预治疗，这样可以降低社会的犯罪率，给社会管理带来积极影响。

六、总 结

本文提出了基于贝叶斯方法研究比例数据分位数回归建模及其统计推断方法。首先借鉴 Tobit 模

型建立了比例数据分位数回归的贝叶斯层次模型，通过选取先验分布提出了 Gibbs 抽样程序。本文提出的方法既可以处理连续型比例数据，又可以处理半连续型比例数据。从模拟研究和实例分析可以看出，我们的估计方法是相当不错的，并且能动态地捕捉因变量与解释变量之间的动态关系，为问题分析提供更丰富的分析结果。在此研究基础上，今后进一步的研究方向是当比例因变量与解释变量之间存在非线性关系时，我们需要研究比例数据分位数回归非参数、半参数建模方法，并研究如何提出有效、快速的估计算法。

香港中文大学 Xinyuan Song 教授和美国加州大学洛杉矶分校 Yih－Ing Hser 教授提供了海洛因吸毒数据，在此特表谢意。

参考文献：

[1] Peter S, Tan M. Marginal Models for Longitudinal Continuous Proportional Data[J]. Biometrics, 2000(56).

[2] Ferrari S, Cribari-Neto F. Beta Regression for Modelling Rates and Proportions[J]. Journal of Applied Statistics, 2004(7).

[3] Papke L, Wooldridge J. Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates[J]. Journal of Applied Econometrics, 1996, 11(6).

[4] Kieschnick R, McCullough B. Regression Analysis of Variates Observed on (0, 1): Percentages, Proportions and Fractions[J]. Statistical Modelling, 2003(3).

[5] Zhao W, Zhang R, Lv Y, Liu J. Variable Selection for Varying Dispersion Beta Regression Model[J]. Journal of Applied Statistics, 2014, 41(1).

[6] Ramalho E, Ramalho J, Murteira J. Alternative Estimating and Testing Empirical Strategies for Fractional Regression Model[J]. Journal Econometrics Survery, 2011, 25(1).

[7] 李泽安，葛建芳，章亚娟. Beta 回归模型在数据挖掘预测中的应用[J]. 南通大学学报：自然科学版，2009，8(3).

[8] 赵为华，张日权. Beta-Binomial 回归模型及其应用[J]. 统计与信息论坛，2016，31(3).

[9] Koenker, Roger Bassett, Gilbert. Regression Quantiles[J]. Econometrica: Journal of the Econometric Society, 1978(1).

[10] Koenker, Roger. Quantile Regression[M]. Cambridge: Cambridge University Press, 2005.

[11] Tobin J. Estimation of Relationships for Limited Dependent Variables[J]. Econometrica, 1958, 26(1).

[12] Amemiya T. Tobit Models: A Survey[J]. Journal of Econometics, 1984, 24(1).

[13] Rahim Alhamzawi, Keming Yu. Bayesian Tobit Quantile Regression Using G-prior Distribution with Ridge Parameter [J]. Journal of Statistical Computation and Simulation, 2015, 85(14).

[14] Kaifeng Zhao, Heng Lian. Bayesian Tobit Quantile Regression with Single-index Models[J]. Journal of Statistical Computation and Simulation, 2015, 85(6).

[15] Andrews D, Mallows C. Scale Mixtures of Normal Distributions[J]. Journal of the Royal Statistical Society, 1974, 36(1).

Bayesian Inference for Quantile Regression of Proportional Data and Its Application

ZHAO Wei-hua[1], ZHANG Ri-quan[2]

(1. School of Science, Nantong University, Nantong 226019, China;
2. School of Statistics, East China Normal University, Shanghai 200241, China)

Abstract: In this paper, we try to use Bayesian method to investigate the regression modeling of the proportional data in the framework of quantile regression. We first give the proposed quantile regression for proportional data based on Tobit model, and then obtain the Bayesian hierarchical model through choosing appropriate prior distributions, which lead to the posterior distribution for Gibbs sampling method. The usefulness and good performance of our proposed method is examined by the simulation studies. Finally, we apply newly proposed method to the heroin use data in California, and reveal the influence factors of drug use frequency at different quantile levels.

Key words: proportional data; quantile regression; Tobit model; Bayesian analysis; Gibbs sampling

（责任编辑：李　勤）