Statistics and Application 统计学与应用, 2021, 10(5), 929-939
Published Online October 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.105097

Hans汉斯

Expectile回归模型的贝叶斯统计推断研究

金　娇*，管思晴*，扈灵嫣*，张凯琦*，吕　洋

北京工业大学，北京

收稿日期：2021年9月30日；录用日期：2021年10月15日；发布日期：2021年10月29日

摘　要

目前，对于回归模型的主要关注点在于，通过描述响应变量分布更一般的特性，扩展具有可行性的(均值)回归模型。本文将利用expectile函数为基础，在贝叶斯的框架下，假设先验函数为正态分布，非对称广义高斯分布为似然分布，建立贝叶斯expectile回归模型，推导了贝叶斯expectile回归的估计方法。本文利用R语言对模型进行了数据模拟和实证分析，利用Metropolis-Hastings算法对目标后验函数进行抽样，从而得出参数的估计，结果表示贝叶斯expectile回归模型具有一定的可行性和准确性。本文还将贝叶斯expectile回归模型应用于美国人口调查工资数据(Berndt, 1991)，对数据进行了回归拟合并计算出置信区间，结果表明，职业、受教育年限等因素对工资存在显著的影响。

关键词

贝叶斯，expectile，非对称广义高斯分布，Metropolis-Hastings算法，R语言

Bayesian Statistical Inference of Expectile Regression Model

Jiao Jin*, Siqing Guan*, Lingyan Hu*, Kaiqi Zhang*, Yang Lv

Beijing University of Technology, Beijing

Received: Sep. 30th, 2021; accepted: Oct. 15th, 2021; published: Oct. 29th, 2021

Abstract

Currently, the main focus for regression models is to extend available (mean) regression models by describing more general properties of the distribution of the response variables. Under the framework of Bayes, based on the expectil, this paper builds the Bayesian expectile regression model which assumes that the prior function is normal distribution and the asymmetric generalized Gaussian distribution is likelihood distribution. Also, this paper deduces the estimation method of Bayesian expectile regression. This paper uses R language for data simulation and empiri-

*共同第一作者。

文章引用: 金娇, 管思晴, 扈灵嫣, 张凯琦, 吕洋. Expectile 回归模型的贝叶斯统计推断研究[J]. 统计学与应用, 2021, 10(5): 929-939. DOI: 10.12677/sa.2021.105097

cal analysis of the model where the Metropolis-Hastings algorithm is used to sample the target posterior function, so as to obtain the parameter estimation. The results show that the Bayesian expectile regression model has certain feasibility and accuracy. In this paper, the Bayesian expectile regression model is applied to the salary data of the United States Population Survey (Berndt, 1991), the confidence interval is calculated and the data is fitted by the regression. The results show that occupation, years of education and other factors have a significant impact on the salary.

Keywords

Bayes, expectile, Asymmetric Generalized Gaussian Distribution, Metropolis-Hastings Algorithm, R Language

1. 引言

贝叶斯统计方法广泛应用于不同的领域，与很多已知的方法进行结合可以得到更好的模型，并不断对新模型进行改进和优化，例如，将贝叶斯方法和 quantile 回归进行结合(Yu 和 Moyeed (2001) [1])。本文考虑将贝叶斯方法与 expectile 模型进行结合，由于抽样具有一定的难度，以及应用不够广泛，学者们至今对于贝叶斯和 expectile 回归结合的研究相对不多。抽样是贝叶斯统计学中重要的环节，由于目标函数(多为后验函数)复杂且不一定存在具体表达式，所以如何能够准确地从贝叶斯目标函数中抽样，是一直以来困扰学者们的问题。近年来，Markov 链蒙特卡罗(MCMC)方法是相对比较主流简单的从后验分布中抽样的方法，其原理是为了对目标函数进行抽样，构建一个可以收敛到目标函数的 Markov 链的稳定分布，通过从这一稳定分布中抽样进而实现从目标函数中抽样。MCMC 方法包括 Metropolis-Hastings 算法和 Gibbs 算法等。本文选择 Metropolis-Hastings 算法对后验函数进行抽样。

2. 数据来源与变量解释

利用贝叶斯 expectile 回归模型对 1985 年 5 月美国人口普查局对当前人口调查(为 1991 年 Berndt 随机抽取样本) [2]。

该数据集包含 11 个变量的 534 个观察值，用于研究工资的决定因素。解释变量包括年龄、潜在工作经验年限、教育程度、居住地区、性别、种族、部门、联盟、职业和婚姻状况信息。由于潜在工作年限 = 年龄 − 受教育年限 − 6，为减少变量的共线性，不考虑年龄作为解释变量。解释变量中受教育年限和潜在工作经验年限为数值型变量，其余为分类变量，其中种族、职业、部门为多分类变量。对于多分类变量，将其划分为 0~1 变量：当 $X_6=0$，$X_7=0$ 时，种族为“cauc”；当 $X_8=0$，$X_9=0$，$X_{10}=0$，$X_{11}=0$，$X_{12}=0$ 时，职业为商人或装配线工人；当 $X_{13}=0$，$X_{14}=0$ 时，部门为其他部门。具体变量解释见表 1：

Table 1. Variable interpretation table
表 1. 变量解释表

变量类型	变量名称	变量描述
因变量	Y	工资，美元/小时

Continued

解释变量	X_1	受教育年限
	X_2	地区，1 = 南方地区，0 = 其他地区
	X_3	性别，1 = 女性，0 = 男性
	X_4	潜在工作经验年限
	X_5	是否在联盟工作，1 = 是，0 = 否
	X_6	种族，1 = 其他种族，0 = 其他
	X_7	种族，1 = 西班牙种族，0 = 其他
	X_8	职业，1 = 管理和行政，0 = 其他
	X_9	职业，1 = 销售员，0 = 其他
	X_{10}	职业，1 = 办公室文员，0 = 其他
	X_{11}	职业，1 = 服务人员，0 = 其他
	X_{12}	职业，1 = 技术或专业工人，0 = 其他
	X_{13}	部门，1 = 制造业或采矿业，0 = 其他
	X_{14}	部门，1 = 建设业，0 = 其他
	X_{15}	是否结婚，1 = 是，0 = 否

3. 研究方法

3.1. 贝叶斯 expectile 回归

建立贝叶斯 expectile 回归方程，假设回归方程为多元线性回归方程，表达式为：

$$y_j = x_j\beta + \varepsilon_j, \tag{1}$$

其中样本数据容量为 n，$x_j = \left(x_{j1}, x_{j2}, \cdots, x_{jp}\right)$，$\beta$ 为未知参数，$\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^{\mathrm{T}}$。误差项 ε_j 服从均值为 0 的分布且独立同分布。将模型用矩阵的形式表达：

$$Y = X\beta + \varepsilon \tag{2}$$

其中 $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$，$Y = \left(y_1, y_2, \cdots, y_n\right)^{\mathrm{T}}$，$\varepsilon = \left(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\right)^{\mathrm{T}}$，$\sigma^2 > 0$ 未知。$x_{j\cdot}$ 表示第 j 次观测自变量的观测值；y_j 表示在自变量 $x_{j\cdot}$ 下因变量第 j 次的观测值。

除了一些众所周知的英文缩写，如 IP、CPU、FDA，所有的英文缩写在文中第一次出现时都应该给出其全称。文章标题中尽量避免使用生僻的英文缩写。

3.1.1. 先验函数

假设先验函数服从正态分布，即 $\pi(\beta) \sim N\left(\mu_\beta, \Sigma_\beta\right)$，其中 $\Sigma_\beta > 0$ 为常数，其概率密度函数为：

$$\pi(\beta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\left(\beta - \mu_\beta\right)^{\mathrm{T}} \Sigma_\beta^{-1} \left(\beta - \mu_\beta\right)\right\} \tag{3}$$

3.1.2. 似然函数

非对称广义高斯分布概率密度函数完全由参数 α，σ_a，σ_b 决定，设 $\alpha = 2$，$\sigma_a = \frac{1}{\sqrt{1-\tau}}$，$\sigma_b = \frac{1}{\sqrt{\tau}}$，

取均值 $\mu=0$， ε 的概率密度函数为：

$$f_\tau(\varepsilon)=\begin{cases}\dfrac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\exp-\left(\dfrac{-\varepsilon}{\sqrt{1-\tau}}\right)^2, & \varepsilon<0\\ \dfrac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\exp-\left(\dfrac{\varepsilon}{\sqrt{\tau}}\right)^2, & \varepsilon\geq 0\end{cases} \tag{4}$$

通过 expectile 函数的损失函数 $\rho_\tau(u)=\left|\tau-I(u<0)\right|u^2$，可以将式(4)化为：

$$f_\tau(\varepsilon)=\left\{\frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\exp\left\{-\rho_\tau(\varepsilon)\right\}\right.$$

那么 $\varepsilon_j, j=1,2,\cdots,n$ 的联合概率密度为：

$$\begin{aligned}L&=\prod_{j=1}^{n}f_\tau(\varepsilon_j)=\prod_{j=1}^{n}f_\tau(y_j-x_j\beta)\\&=\left[\frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\right]^n\cdot\exp\left\{-\sum_{\varepsilon_j<0}\left(\frac{-\varepsilon_j}{\sqrt{1-\tau}}\right)^2-\sum_{\varepsilon_j\geq 0}\left(\frac{\varepsilon_j}{\sqrt{\tau}}\right)^2\right\}\\&=\left[\frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\right]^n\cdot\exp\left\{-\sum_{j=1}^{n}\rho_\tau(\varepsilon_j)\right\}\end{aligned} \tag{5}$$

假设因变量 y_j 服从非对称广义高斯分布，则似然函数根据式(5)可以写为：

$$L(Y\mid\beta)=\left[\frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\right]^n\cdot\exp\left\{-\sum_{j=1}^{n}\rho_\tau(y_j-x_j\beta)\right\} \tag{6}$$

另 $\varepsilon_j=y_j-x_j\beta$ 表示残差，其中 τ 是固定值， β 是未知参数，需要通过抽样得到。

3.1.3. 先验函数

参数 β 的后验函数分布密度与先验分布密度和似然函数的乘积成正比，表达式为：

$$\pi(\beta\mid Y)\propto L(Y\mid\beta)\pi(\beta)$$

由式(3)和式(6)可得：

$$\begin{aligned}\pi(\beta\mid X,Y)&\propto\exp\left\{-\sum_{j=1}^{n}\rho_\tau(y_j-x_j\beta)\right\}\cdot\exp\left\{-(\beta-\mu_\beta)^{\mathrm{T}}\Sigma_\beta^{-1}(\beta-\mu_\beta)\right\}\\&\propto\exp\left\{-\sum_{j=1}^{n}\rho_\tau(y_j-x_j\beta)-(\beta-\mu_\beta)^{\mathrm{T}}\Sigma_\beta^{-1}(\beta-\mu_\beta)\right\}\end{aligned} \tag{7}$$

3.2. Metropolis-Hastings 算法

由式(7)可知，后验分布没有具体的表达式，因此不能用 MCMC 法进行直接抽样，由于 Metropolis-Hastings 算法适用于只知道后验分布正比于先验分布和似然函数的情况，因此本文选择 Metropolis-Hastings 算法进行样本抽样。Metropolis-Hastings 方法是一般化的 MCMC 方法，它由 Metropolis

等人在 1953 年[3]提出了一种构造转移方法，Hastings 随后对这个算法进行了进一步的扩展[4]，形成了 Metropolis-Hastings 方法。

假设建议随机变量 $W' \sim q(\omega' \mid \omega)$ 且通过 $W_t' = W^{(t)} + K_t$ 产生，其中 K_t 与 $W^{(t)}$ 独立，成为随即增量，服从某个关于 0 对称的固定的随机分布。此时的抽样方法称为随机游动 MH 抽样，建议分布 $q(\omega' \mid \omega) = q(\omega' - \omega)$。

利用 Metropolis 选择进行建议分布的选择，考虑对称的建议分布

$$q(\omega \mid \omega') = q(\omega' \mid \omega)$$

此时的接受率简化为：

$$\alpha(\omega, \omega') = \min\left\{1, \frac{\pi(\omega' \mid x)}{\pi(\omega \mid x)}\right\} \tag{8}$$

因此建议分布为 $q(\omega' \mid \omega) = q(|\omega - \omega'|)$。该方法的函数形式简单，抽样的过程直接，同时算法收敛，因其随机游走链趋向于高后验概率密度的区域。

对于贝叶斯 expectile 回归模型，式(8)为：

$$\alpha(\beta, \beta') = \min\left\{1, \frac{\pi(\beta' \mid Y)}{\pi(\beta \mid Y)}\right\} = \min\left\{1, \frac{\pi(\beta')L(Y \mid \beta')}{\pi(\beta)L(Y \mid \beta)}\right\}$$

其中 $\pi(\cdot)$ 表示先验函数，$L(\,)$ 表示似然函数，假设先验分布服从正态分布，即 $\pi(\beta) \sim N(\mu_\beta, \Sigma_\beta)$。假设 u 服从均匀分布即 $u \sim U(0,1)$，则

$$\beta^{(t+1)} = \begin{cases} \beta', u \le \alpha(\beta, \beta') \\ \beta, u > \alpha(\beta, \beta') \end{cases}$$

$\beta_t' = \beta^{(t)} + K_t$，假设 $K_t \sim N(0, cI_n)$，那么 $\beta_t' \sim N(\mu_{\beta^{(t)}}, cI_n)$，其中取均值的初始值为最小二乘估计值:

$$\beta^{(0)} = \arg\min_\beta \sum_{j=1}^{n} \left(y_j - x_j^{\mathrm{T}}\beta\right)^2$$

K_t 的取值决定了建议分布选择的准确性，K_t 是每一步游走的步长，不当的选择会减缓抽样速度以及导致接受率过大或过小。当接受概率过小时，随机游走几乎不会移动，要想随机游走整个可行域需要非常大的抽样次数，收敛范围不理想，容易产生一定的误差；当接受概率过大时，更新值接近于抽样值，随机游走步长过小，抽样次数同样需要很大。因此目标为选择合适的抽样次数和 c 来保证合适的接受率。

通常情况下，接受率在 0.4~0.6 之间为最佳，为保证抽样结果收敛，本文选择 n 次迭代并取后 $n/2$ 次迭代结果的均值来估计参数 β。

$$\bar{g}(\beta) = \frac{1}{n/2}\sum_{t=n/2+1}^{n} g\left(\beta^{(t)}\right)$$

4. 模型建立

4.1. 模拟数据

为了更全面的验证贝叶斯 expectile 回归模型的可行性和精度，利用 R 语言对不同样本容量，不同 τ，不同分布下的随机误差产生样本数据，从而更准确的比较不同样本数据下贝叶斯 expectile 回归模型对于参数 β 的估计效果。

模拟数据的设计如下：另模拟的数据模型为：

$$y_j = x_j\beta + \varepsilon_j$$

其中 β 服从多元高斯正态分布，为待估计的参数，ε_j 为独立同分布的残差项。为避免真值不同对估计产生的误差，设定相同的真值，相同的迭代次数；为避免随机游动 Metropolis-Hastings 算法中游走步长对估计产生的误差，选择固定的步长。根据估计值与真实值的误差和均方误差进行模型的准确性比较。

假设真值 $\beta=(1.2,-1.35,0,-0.7)^{\mathrm{T}}$，迭代次数 $B=1000$，游走步长 $K_t \sim N(0,cI_n)$ 中 c 固定，I_n 是单位矩阵。

1) 假设随机误差服从标准正态分布，即 $\varepsilon_j \sim N(0,1)$，$\tau=0.5$，模拟数据研究不同样本容量 n 下模型的估计值，估计值与真实值的误差和均方误差，以及模型的接受率见表 2：

Table 2. Model comparison under different sample sizes
表 2. 不同样本容量下的模型比较

	n = 100		n = 300		n = 500	
接受率	0.704		0.565		0.464	
	偏差	MSE	偏差	MSE	偏差	MSE
β_1	0.002186	0.121372	−0.001465	0.069440	0.001543	0.053608
β_2	0.000469	0.146966	−0.000642	0.077224	−0.001522	0.060419
β_3	−0.000153	0.142709	−0.000771	0.075833	−0.001034	0.059508
β_4	−0.004215	0.118435	0.000095	0.069156	0.001937	0.051630

由表 2 可以看出，在固定随机误差分布和 τ 时，由于估计的偏差都在 0 附近且很小，MSE 也很小。因此可以判断利用贝叶斯 expectile 模型估计得到的参数 β 的估计值是比较稳定的，是比较趋于真实值的。对于 $\beta_i, i=1,2,3,4$，随着样本量的增加，MSE 减小，这说明模型有效，模型的估计值更准确。对于样本容量为 100 的时候，接受率为 0.704，接受率偏大。当接受率偏大时，需要更多的抽样次数才可以保证模型的准确性。

假设随机误差服从自由度为 4 的 t 分布，即 $\varepsilon_j \sim t(4)$，样本容量 n = 300，模拟数据研究不同 τ 下模型的估计值，估计值与真实值的误差和均方误差，以及模型的接受率见表 3：

Table 3. Comparison of models with different τ
表 3. 不同 τ 下的模型比较

ε_j 分布	$\tau=0.2$		$\tau=0.4$		$\tau=0.6$		$\tau=0.8$	
接受率	0.576		0.532		0.534		0.546	
	偏差	MSE	偏差	MSE	偏差	MSE	偏差	MSE
β_1	0.002982	0.115111	−0.003235	0.099001	−0.0032	0.098177	0.000667	0.110123
β_2	−0.001715	0.12274	0.003186	0.11241	0.001387	0.111841	−0.00614	0.129831
β_3	0.001022	0.12625	−0.002347	0.112426	0.002387	0.114051	0.006281	0.127938
β_4	0.000880	0.109788	0.001134	0.094887	−0.005973	0.096936	−0.001387	0.114802

由表 3 可以看出，在固定随机误差分布和样本容量时，估计的偏差都在 0 附近且很小，MSE 也很小。因此可以判断利用贝叶斯 expectile 模型估计得到的参数 β 的估计值是比较稳定的，是比较趋于真实值的。

不同 τ 下的 MSE 相差不大，当 $\tau=0.2$ 时，参数估计的偏差最小，因此该情况下的模型估计相对更为准确。接受率处于 0.4~0.6 之间，为合适接受率，可以判断 τ 对模型的估计存在一定的影响。

假设样本容量 $n=500$，$\tau=0.8$，研究随机误差服从标准正态分布，t 分布和非对称广义高斯分布下模型的估计值，估计值与真实值的误差和均方误差，以及模型的接受率见表 4：

Table 4. Model comparison under different random error distributions
表 4. 不同随机误差分布下的模型比较

ε_j 分布	$N(0,1)$		t (4)		AGGD	
接受率	0.476		0.565		0.364	
	偏差	MSE	偏差	MSE	偏差	MSE
β_1	0.001179	0.061503	0.000667	0.110123	−0.000276	0.016644
β_2	0.0014126	0.069806	−0.00614	0.129831	−0.000350	0.019154
β_3	0.002543	0.070753	0.006281	0.127938	−0.000022	0.018804
β_4	0.000102	0.061798	−0.001387	0.114802	−0.000322	0.016117

其中 $\varepsilon_j \sim AGGD$ 的密度函数为

$$f_\tau(\varepsilon_j)=\begin{cases}\dfrac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\exp-\left(\dfrac{-\varepsilon_j}{\sqrt{1-\tau}}\right)^2, & \varepsilon_j<0\\ \dfrac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi}\left(\sqrt{\tau}+\sqrt{1-\tau}\right)}\exp-\left(\dfrac{\varepsilon_j}{\sqrt{\tau}}\right)^2, & \varepsilon_j\geq 0\end{cases}$$

由表 4 可以看出，在固定 τ 和样本容量时，估计的偏差都在 0 附近且很小，MSE 也很小。因此可以判断利用贝叶斯 expectile 模型估计得到的参数 β 的估计值是比较稳定的，是比较趋于真实值的。不同随机误差分布下，参数的偏差和 MSE 存在一定差异，可以看出当随机误差服从非对称广义高斯分布时，偏差和 MSE 最小，因此该情况下的模型估计相对更为准确。当随机误差服从标准正态分布和 t 分布时，接受率处于 0.4~0.6 之间，为合适接受率；随机误差服从非对称广义高斯分布时，接受率小于略小于 0.4，由 3.3 结论可以知道，应该增加样本容量以增加模型准确性。

因此可以得出结论，随机误差的分布对模型的估计存在较为显著的影响，当随机误差服从非对称广义高斯分布时，模型的准确性更高。

4.2. 实际数据分析

4.2.1. 数据来源与变量解释

该数据集包含 11 个变量的 534 个观察值，用于研究工资的决定因素。解释变量包括年龄、潜在工作经验年限、教育程度、居住地区、性别、种族、部门、联盟、职业和婚姻状况信息。由于潜在工作年限 = 年龄 − 受教育年限 −6，为减少变量的共线性，不考虑年龄作为解释变量。解释变量中受教育年限和潜在工作经验年限为数值型变量，其余为分类变量，其中种族、职业、部门为多分类变量。对于多分类变量，将其划分为 0~1 变量：当 $X_6=0$，$X_7=0$ 时，种族为“cauc”；当 $X_8=0$，$X_9=0$，$X_{10}=0$，$X_{11}=0$，$X_{12}=0$ 时，职业为商人或装配线工人；当 $X_{13}=0$，$X_{14}=0$ 时，部门为其他部门。具体变量解释见表 5：

Table 5. Variable explanation table
表 5. 变量解释表

变量类型	变量名称	变量描述
因变量	Y	工资，美元/小时
解释变量	X_1	受教育年限
	X_2	地区，1 = 南方地区，0 = 其他地区
	X_3	性别，1 = 女性，0 = 男性
	X_4	潜在工作经验年限
	X_5	是否在联盟工作，1 = 是，0 = 否
	X_6	种族，1 = 其他种族，0 = 其他
	X_7	种族，1 = 西班牙种族，0 = 其他
	X_8	职业，1 = 管理和行政，0 = 其他
	X_9	职业，1 = 销售员，0 = 其他
	X_{10}	职业，1 = 办公室文员，0 = 其他
	X_{11}	职业，1 = 服务人员，0 = 其他
	X_{12}	职业，1 = 技术或专业工人，0 = 其他
	X_{13}	部门，1 = 制造业或采矿业，0 = 其他
	X_{14}	部门，1 = 建设业，0 = 其他
	X_{15}	是否结婚，1 = 是，0 = 否

4.2.2. 贝叶斯 expectile 回归模型

考虑因变量和自变量服从贝叶斯 expectile 回归模型，

$$y_j = x_j\beta + \varepsilon_j,\ j = 1,2,3,\cdots,534$$

其中 y_j 表示工资，x_j 表示这个人的各项影响因素，β 为未知待估参数，ε_j 是误差项。由数据模拟可知，ε_j 独立同分布于非对称广义高斯分布时，模型的准确性更高。其中 expectile 函数的损失函数为：

$$\rho_\tau(u) = \left|\tau - I(u<0)\right|u^2$$

由数据模拟的结论，取 $\tau = 0.2$。

选择先验分布为正态分布，即 $\pi(\beta) \sim N(\mu_\beta, \Sigma_\beta)$，由式(3-8)可知后验分布函数为

$$\pi(\beta \mid X, Y) \propto \exp\left\{-\sum_{j=1}^{534}\rho_\tau\left(y_j - x_j\beta\right) - \left(\beta-\mu_\beta\right)^{\mathrm{T}}\Sigma_\beta^{-1}\left(\beta-\mu_\beta\right)\right\}$$

利用随机游动 MH 抽样对数据进行抽样，其中取均值的初始值为最小二乘估计值：

$$\beta^{(0)} = \arg\min_\beta \sum_{j=1}^{534}\left(y_j - x_j^{\mathrm{T}}\beta\right)^2$$

迭代 10,000 次，取后 5000 次的平均值为贝叶斯 expectile 回归模型的参数估计值。不断调整 c 使得其接受率处于 0.4~0.6 之间。

4.2.3. 回归结果

经过不断调整后，随机游动 MH 抽样的接受率等于 0.4103，参数 β 的最小二乘估计值、贝叶斯 expectile 回归函数估计及其 95%置信区间的数据可以在表 6 中找到：

Table 6. Parameter estimation table
表 6. 参数估计表

变量名称	最小二乘估计	贝叶斯 expectile	置信区间
X_1	0.58402107	0.58175824	[0.5838222, 0.58421993]
X_2	−0.72647934	−0.62788149	[−0.7279338, −0.72502488]
X_3	−2.07362509	−1.97556174	[−2.0745829, −2.07266727]
X_4	0.07942931	0.08051625	[0.0793186, 0.07954001]
X_5	1.76631319	1.59732956	[1.7644343, 1.76819210]
X_6	−0.82105112	−0.85605366	[−0.8222906, −0.81981159]
X_7	−0.74089581	−0.74268284	[−0.7446778, −0.73711382]
X_8	3.50394788	3.41874156	[3.5025716, 3.50532420]
X_9	−0.78898500	−0.73434984	[0.7905009, −0.78746908]
X_{10}	0.17105910	0.04537268	[0.1693982, 0.17271997]
X_{11}	−0.53868936	−0.75394858	[−0.5412502, −0.53612849]
X_{12}	2.30651389	2.15683655	[2.3048609, 2.30816690]
X_{13}	1.10343959	0.97557239	[1.1020718, 1.10480735]
X_{14}	0.32834338	0.37741230	[0.3267033, 0.32998350]
X_{15}	0.32946753	0.25858624	[0.3277690, 0.33116606]

根据结果可知，贝叶斯 expectile 回归模型的估计值与初值非常接近，模型拟合度良好。对于一个给定的贝叶斯 95%可信区间，其意义是参数真值有 95%的概率落在这个区间内。根据参数 β 的置信区间，若置信区间包括 0，则变量不显著。可以看出，各变量的置信区间都不包括 0，这说明了潜在工作经验年限、教育程度、居住地区、性别、种族、工作部门、联盟工作、职业和婚姻状况都对工资有着显著的影响。

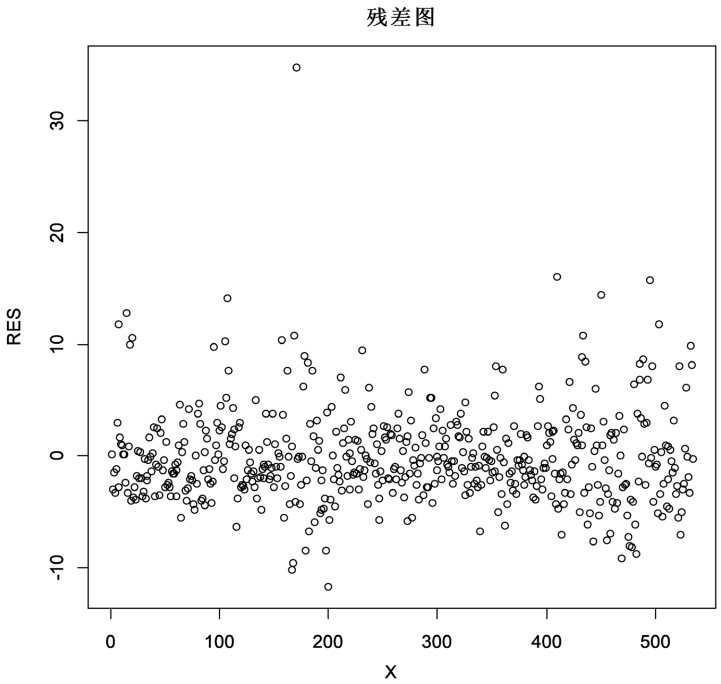

Figure 1. Residual diagram
图 1. 残差图

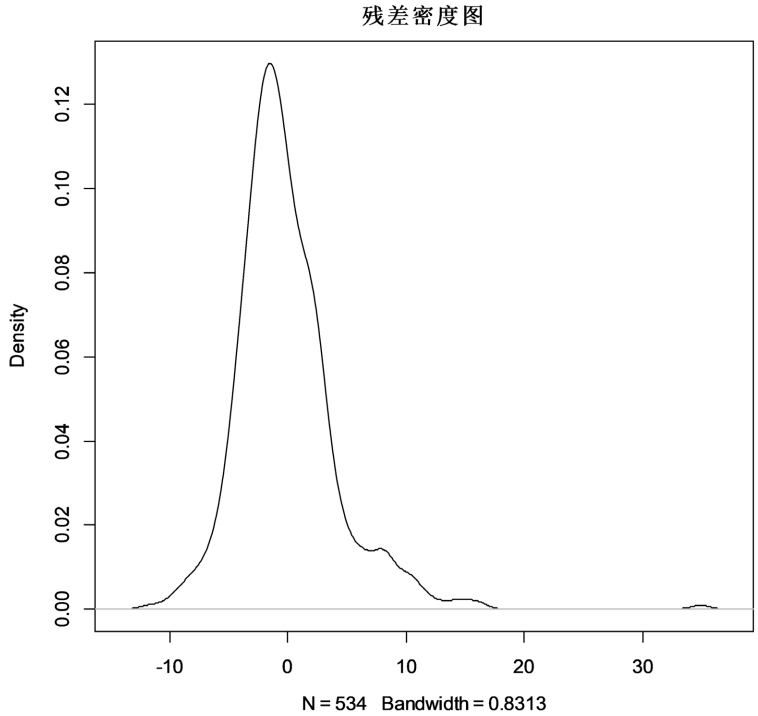

Figure 2. Residual density diagram
图 2. 残差密度图

由图 1 的残差图可知，残差呈随机分布，无明显漏斗状，因此认为残差是等方差的。由图 2 残差密度图可以看出，残差近似于关于 0 对称的正态分布，这与假设残差服从均值为 0 的非对称广义高斯分布相吻合。由于残差越接近于 0，模型的拟合度越高，模型越准确，参数的估计越精确。因此，由残差图和残差密度图可以判断出，大多数残差接近于 0，贝叶斯 expectile 回归模型的准确率较高。

根据估计系数可以得出以下结论：

1) 受教育年限和潜在的工作年限与工资(美元/小时)成正比关系；

2) 南方地区工资比其他地区的工资低 0.627 (美元/小时)；

3) 女性比男性的工资低 1.975 (美元/小时)；

4) 在联盟工作的人工资比不在联盟工作的人高 1.597 (美元/小时)；

5) 对于种族而言，其他种族的工资比种族属于“cauc”和西班牙的工资低 0.856 (美元/小时)；属于西班牙种族的人均工资比其他种族和“cauc”的工资低 0.742 (美元/小时)；

6) 对于职业而言，行政和管理人员的工资高于其他职业工资 3.418 (美元/小时)；销售人员工资低于其他职业工资 0.734 (美元/小时)；办公室文员的工资高于其他职业工资 0.045 (美元/小时)；服务人员工资低于其他职业工资 0.753 (美元/小时)；技术或专业工人工资高于其他职业工资 2.156 (美元/小时)；

7) 对于工作部门而言，制造业或采矿业工资高于其他职业工资 0.975 (美元/小时)；建设业工资高于其他职业工资 0.377 (美元/小时)；

8) 已婚人士工资比未婚人士工资高 0.258 (美元/小时)。

综上，贝叶斯 expectile 回归模型拟合度较好，参数估计的准确率较高。其中受教育年限越长，在联盟工作，潜在工作年限越长，工作部门属于制造业或采矿业，职位是行政和管理人员，种族属于“cauc”的其他地区已婚男性，相对拥有更高的工资。

5. 模型评价

1) 该方法用贝叶斯 expectile 回归模型进行了数据模拟和实证分析，有良好的可行性和准确性。

2) 在不同 τ 和不同随机误差分布的情况下，对模型进行数据模拟，可以得出模型良好可行性。

3) 通过计算估计参数的偏差和均方误差来评估模型，发现该模型有较强准确性。

4) 对于分类变量，给出了同因素下不同类别的差异比较；对于连续变量，给出了与工资的相关关系。可对真实情况有一定的参考。

6. 结论

本文结合了贝叶斯统计和 expectile，提出了贝叶斯 expectile 回归模型。文章第三章和第四章从理论和数据模拟与实证研究，研究了该模型的可行性和准确性。具体来说，本文第三章利用正态分布作为先验函数，非对称广义高斯分布作为样本的似然函数，在 Metropolis-Hastings 抽样方法的背景下，对模型进行理论推导，给出了贝叶斯 expectile 的后验函数以及贝叶斯 expectile 回归模型的参数估计计算表达式。本文第四章从模拟数据和实证分析，证明了贝叶斯 expectile 回归模型给出的估计值与真实值误差较小。对于模拟数据，在真值、迭代次数与随机游走步长相同的情况下，且在不同样本容量、 τ 和随机误差分布的情况下，对模型进行数据模拟，可以得出模型具有可行性，通过计算估计参数的偏差和均方误差来估计模型的准确性。当随机误差服从 t 分布，样本容量固定的情况下， $\tau = 0.2$ 时，模型参数的估计更准确；当样本容量固定， τ 固定的情况下，随机误差服从非对称广义高斯分布时，模型参数的估计更准确；当 τ 固定的情况下，随机误差服从标准正态分布时，样本容量越高，模型参数的估计更准确。对于实证分析，利用 1985 年 5 月美国人口普查局对当前人口调查的抽样结果，研究不同因素对工资的影响，对于连续变量，受教育年限和潜在工作年限与工资成正比；对于分类变量，在联盟工作，工作部门属于制造业或采矿业，职位是行政和管理人员，种族属于“cauc”，其他地区，已婚，男性，相对拥有更高的工资。

参考文献

[1] Yu, K. and Moyeed, R.A. (2001) Bayesian Quantile Regression. *Statistics & Probability Letters*, **54**, 437-447. https://doi.org/10.1016/S0167-7152(01)00124-9

[2] Berent, E.R. (1992) The Practice of Econometrics: Classical and Contemporary. Addison Wesley, Reading, MA.

[3] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., *et al*. (1953) Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21**, 1087-1092. https://doi.org/10.1063/1.1699114

[4] Hastings, W.K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109. https://doi.org/10.1093/biomet/57.1.97