贝叶斯统计和经典统计在分位数回归分析中的比较

谷恒明[1],胡良平[1,2]

[摘要] **目的** 在分位数回归分析中比较贝叶斯统计和经典统计,以便在不同场合下选择更加有效的方法。**方法** 选择大样本数据,基于经典统计和贝叶斯统计的分位数回归分析利用 SAS 软件中的 QUANTREG 过程和 MCMC 过程实现。分别采用十折交叉验证方法,通过训练集的拟合效果和预测集的预测效果两方面来评价模型优劣。**结果** 若采用全部样本建立模型时,基于经典统计的分位数回归分析评价指标略差于基于贝叶斯统计的分位数回归分析评价指标;基于部分样本作为训练集的十折交叉验证时,比较 10 次指标的均值,基于贝叶斯统计相对于基于经典统计而言,在具体的分位数回归方程中,其下四分位数(Q_1)和上四分位数(Q_3)的拟合效果为优,而中位数(Q_2)的拟合效果略差;对于预测效果而言,基于贝叶斯统计的分位数回归方程要优于经典统计的分位数回归方程。**结论** 在拟解决实际问题的场合下,如要求准确度较高,主要考察各个分位数预测效果和拟合效果,可选择贝叶斯分位数回归分析法;若主要考察中位数的拟合效果则需要谨慎选择。如时间精力有限且样本量足够大,那么采用经典统计的分位数回归分析即可。

[关键词] 贝叶斯统计;经典统计;分位数回归;拟合效果;预测效果;交叉验证
[中图分类号] C8 [文献标志码] A
[文章编号] 1674-9960(2018)02-0149-05 **DOI**:10.7644/j.issn.1674-9960.2018.02.014

Comparison of Bayesian statistics and classical statistics in quantile regression analysis

GU Heng-ming[1], *HU Liang-ping*[1,2*]
(1. Consulting Center for Biomedical Statistics, Graduate School, Academy of Military Sciences, Beijing 100850, China; 2. Specialty Committee of Clinical Scientific Research Statistics, World Federation of Chinese Medicine Societies, Beijing 100029, China)
* Corresponding author, Tel: 13126501849, E-mail: lphu812@sina.com

[**Abstract**] **Objective** To compare the Bayesian statistics and the classical statistics in the quantile regression analysis in order to select a more effective method. **Methods** The large sample data was chosen, and the QUANTREG procedure in SAS was used for the classical statistics and the MCMC procedure in SAS for the Bayesian one, respectively. Using ten-fold cross-validation method, the goodness of fitting of the models was appraised in terms of the fitted effect based on the training dataset and the predicted effect based on the predictive dataset. **Results** In most cases, the indexes of the quantile regression models in the classical statistics were slightly worse than those of the Bayesian one. In the ten-fold cross-validation of the partial samples as a training dataset, the fitting effect of the lower quartile (Q_1) and upper quartile (Q_3) of the Bayesian statistics was better than that of the classical one. However, the median (Q_2) fitting effect of the Bayesian statistics was slightly worse than that of the classical one. As for the prediction effect, the Bayesian statistical quantile regression model was superior to the classic one. **Conclusion** To expect high accuracy, such as the predictive effects and fitting effects of each quantile, the Bayesian quantile regression analysis should be chosen. If the major concern is the fitting effect of the median, careful selection from the approaches mentioned above is needed. If time and energy are limited, and the sample size is large enough, the classic statistical quantile regression analysis is a good choice.

[**Key words**] Bayesian statistics; classical statistics; quantile regression; fitting effect; prediction effect; cross-validation

经典最小二乘回归分析主要解决因变量与自变量之间的线性关系,描述自变量对因变量均值的影响[1]。在给定自变量的条件下,因变量 Y 的分布严重偏离正态分布时,用基于最小二乘回归分析原理得到的多重线性回归模型来揭示因变量与自变量之间的依赖关系就会大相径庭。1978 年,Koenker 等[2]在最小绝对偏差估计理论的基础上首次提出了分位数回归分析的概念。在给定因变量一个条件

[基金项目] 国家高技术研究发展计划资助项目(2015AA020102)
[作者简介] 谷恒明,男,硕士,研究方向:比较统计学,Tel:15701577074,E-mail:723761414@qq.com
[作者单位] 1.军事科学院研究生院生物医学统计学咨询中心,北京 100850;2.世界中医药学会联合会临床科研统计学专业委员会,北京 100029
[通讯作者] 胡良平,Tel:13126501849,E-mail:lphu812@sina.com

分位数的情况下,分别对自变量进行线性回归,即此时的线性回归分析不再是针对给定自变量的条件下因变量的均值,而是一个特定的百分位数,如 25% 分位数(即四分位数)。分位数回归分析可全面描绘不同百分位数下影响因素的作用情况,因此相对于估计因变量均值的最小二乘回归分析而言能更好地捕捉数据的重要特征。分位数回归虽可解决线性回归的参数估计问题,但研究者也发现分位数回归的局限性,如在研究样本很小时,得出的模型效果很差[3]。本文主要研究贝叶斯统计方法与经典统计方法在分位数回归分析中的优劣情况,因此,有意识地选取具有足够大样本量的实际问题来开展比较研究。

1 基于两种统计思想构建分位数回归模型

1.1 经典统计分位数回归模型

设用 Y 代表因变量,X_1、X_2、……、X_m 分别代表 m 个自变量,Y_i 表示因变量的第 i 百分位数,则分位数回归模型可以表示为:

$$Y_i = \beta_{i0} + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots\cdots + \beta_{im}X_m + \varepsilon_i \quad (1)$$

式中,β_{i0} 为截距,β_{i1}、β_{i2}、……、β_{im} 分别为各个自变量所对应的偏回归系数,ε 为随机误差。偏回归系数 $\beta_{ij}(j=1,2,\cdots\cdots,m)$ 表示在其他自变量固定不变的条件下,X_n 每改变一个测量单位时所引起的因变量 Y_i 的平均改变量。样本回归方程可以表示为:

$$\hat{Y_i} = b_{i0} + b_{i1}X_i + b_{i2}X_2 + \cdots\cdots + b_{im}X_m \quad (2)$$

式中,$\hat{Y_i}$ 表示第 i 百分位数的估计值,b_{i0}、b_{i1}、b_{i2}、……、b_{im} 为截距和偏回归系数的样本估计值[4]。

1.2 贝叶斯回归模型

$$Y_i = \mu_i + \varepsilon_i, \mu_i = \beta_{i0} + \beta_{i1}X_1 + \cdots\cdots + \beta_{im}X_m \quad (3)$$

此处要为各个参数指定一个先验分布,例如:

$$\pi(\beta_0) = \phi(0, var = le6)$$

$$\pi(\beta_1) = \phi(0, var = le6)$$

$$\pi(\varepsilon_i) = f_{i\Gamma}(shape = 3/10, scale = 10/3) \quad (4)$$

式(3)中,变量的含义与式(1)相同,与经典回归分析相比,贝叶斯回归分析要为参数设定一个先验分布。式(4)为设定的先验分布,截距项 β_0 和偏回归系数 $\beta_{ij}(j=1,2,\cdots\cdots,m)$ 服从正态分布 $N\sim(0, 10^6)$,ε_i 服从位置参数为 3/10,尺度参数为 10/3 的均匀分布。经典统计思想假定回归参数固定,而贝叶斯统计思想假定回归参数是随机的[5]。基于贝叶斯统计思想建立回归模型时,要为各个参数(即回归系数)和残差指定各自的先验分布,可依靠经验或预试验的结果指定合适的先验分布;如无法给定先验分布,可使用无信息先验(相当于均匀分布)代替。

如何进行分位数回归分析呢?经典统计建模可利用 SAS 软件的 STAT 模块中 QUANTREG 过程来实现;而贝叶斯统计建模可利用 SAS 软件的 STAT 模块中 MCMC 过程来实现。MCMC 方法,即马氏链蒙特卡罗方法,默认算法是使用正态分布随机游动 Metropolis 算法。

MCMC 的抽样方法有 Gibbs 抽样、Metropolis 抽样、独立性抽样、随机游动 Metropolis 抽样等。

2 基于两种统计思想拟合分位数回归模型效果的评价

2.1 用全部数据来建立模型并比较

评价指标使用相对误差绝对值的平均值(以下简写为 *Abserror*)、决定系数(R^2)和残差平方和(*SSres*)。

2.2 用部分样本建立模型,其余样本用作预测

采用 K-Fold 交叉验证,即全部数据拆分为 K 份,其中 $(K-1)$ 份用作建立模型的训练集,剩余一份作为测试集[6]。训练集拟合效果使用 *Abserror*,R^2 和 *MSE* 作为评价指标;测试集使用 *Abserror*,*MSE* 与标准化均方误差(*NMSE*)作为评价指标。在 K-Fold 交叉验证中,K 取 10,即进行 10 次建模,每次使用其中 90% 数据建立模型。

2.3 评价指标的具体公式

分位数回归分析的决定系数 R^2 与一般线性回归分析计算的 R^2 有所不同。本文中主要进行下四分位数、中位数及上四分位数的回归曲线的比较,因此决定系数有 3 个,分别用 R^2_{25},R^2_{50} 和 R^2_{75} 代表。计算公式如下:

$$R^2_{25} = \frac{\sum(y - y_{25})^2 - \sum(y - \hat{y_{25}})^2}{\sum(y - y_{25})^2} \quad (5)$$

以 25% 分位数为例,y 为原始值,y_{25} 表示下四分位数,$\hat{y_{25}}$ 为预测值。

$$MSE = \frac{\sum(y - \hat{y_{25}})^2}{N - n - 1} \quad (6)$$

同样以 25% 分位数为例,N 为样本量,n 为自变量个数。从上述公式可见,*MSE* 是残差平方和与自由度的比值。交叉验证中 K 取值不同,建立模型的训练集和预测使用的测试集样本量是不同的,直接基于残差平方和的比较不够合理,因此,需要除以自由度。

$$NMSE = \frac{\sum(y - \hat{y_{25}})^2}{\sum(y - y_{25})^2} \quad (7)$$

同样以 25% 分位数为例。在数值上,*NMSE* 等于 $1 - R^2$,R^2 为回归分析的决定系数。但对于测试集来说,其 *NMSE* 与测试集回归的 R^2 无关。交叉验

证主要关心测试集的 *NMSE*。

由于分位数回归对模型误差项的分布并无要求,因此其回归系数相对于最小二乘回归分析更加稳健。普通的多重线性回归要求因变量满足正态分布,因此只能得出对均值拟合的回归系数[7]。

3 基于全部样本建模,经典统计与贝叶斯统计方法的比较

3.1 经典统计的分位数回归分析

通过 QUANTREG 过程,可得到因变量各个分位数的参数估计。为节省篇幅,本文只采取其中下四分位数、中位数及上四分位数。以 SAS 软件中 Sashelp. bweight 数据中的前 1000 例作为研究对象,比较两者的优劣。

【例】某医学研究课题拟对影响初生婴儿的几项重要因素进行研究,其中因变量 Weight 不满足正态分布,因此,若采用一般的多重线性回归分析是不合理的,这里采用分位数回归分析。变量名称及含义的详细情况见表 1,数据见表 2。

表 1 影响初生婴儿体质量的几项重要因素及其含义

变量名称	含义描述
Weight	初生婴儿体质量
Black	1 表示黑人,0 表示白人
Married	1 表示已婚,0 表示未婚
Boy	1 表示男孩,0 表示女孩
Visit	产前第一次检查:0 表示未作检查,1 表示中间 3 个月,2 表示最后 3 个月,3 表示前 3 个月
Ed	母亲受教育程度:0 表示高中毕业,1 表示上大学但未毕业,2 表示本科,3 表示高中以下
Smoke	1 表示吸烟,0 表示不吸烟
CigsPer	每天吸烟支数
Mom_Age	母亲的年龄
M_WtGain	母亲怀孕期间体质量增值

表 2 1000 例初生婴儿体质量及其影响因素的部分数据

Weight	Black	Married	Boy	Mom_Age	Smoke	Cigsper	M_WtGain	Visit	Ed
4111	0	1	1	-3	0	0	-16	1	0
3997	0	1	0	1	0	0	2	3	2
3572	0	1	1	0	0	0	-3	3	0
1956	0	1	1	-1	0	0	-5	3	2
3515	0	1	1	-6	0	0	-20	3	0
					.....				
3804	1	0	1	-5	0	0	-20	3	1
3459	0	0	0	-1	0	0	-14	1	3
2982	0	0	0	-4	0	0	-5	3	1
3856	0	1	1	-3	0	0	3	3	2
4196	0	0	0	-8	1	20	-5	0	0

初生婴儿体质量 Weight 的四分位数分别为:上四分位数 3714. 0,中位数 3402. 0,下四分位数 3047. 5。

基于经典统计的分位数回归分析所得的四分位数评价结果见表 3。

表 3 基于经典统计的分位数回归分析所得的四分位数评价结果

模型评价指标	*Abserror*	R^2	*SSress*
下四分位数(Q_1)	0. 1526075497	0. 092462	371545836. 34
中位数(Q_2)	0. 1382708677	0. 084081	283465524. 08
上四分位数(Q_3)	0. 1731885146	0. 089527	391049337. 17

由表 3 可以看出,基于经典统计的分位数回归分析所得的四分位数拟合效果并不理想。决定系数 R^2 太小,其中,对中位数拟合的效果稍好。

3.2 贝叶斯统计的分位数回归分析

贝叶斯统计分位数回归分析中的四分位数评价结果见表 4。

表 4 贝叶斯统计分位数回归分析中的四分位数评价结果

模型评价指标	*Abserror*	R^2	*SSress*
下四分位数(Q_1)	0. 1411850026	0. 24689	308321175. 14
中位数(Q_2)	0. 138179576	0. 09051	281475789. 98
上四分位数(Q_3)	0. 1494973045	0. 28161	308550320. 28

由表 4 可见,基于贝叶斯统计的分位数拟合效果虽并不理想,但在与误差有关的评价指标的数值上要小于经典统计的分位数回归分析相应的结果。其中,虽对中位数的拟合残差平均值的绝对值和残差平方和最小,但决定系数 R^2 也是最小的。

4 基于部分样本建模,经典统计与贝叶斯统计方法的比较

4.1 基于经典统计分位数回归分析的训练集拟合效果

基于经典统计分位数回归分析中的拟合效果评价结果见表 5。

4.2 基于经典统计分位数回归分析的测试集的预测效果

基于经典统计分位数回归分析中的预测效果评价结果见表 6。

4.3 基于贝叶斯统计分位数回归分析的训练集的拟合效果

贝叶斯统计分位数回归分析中的训练集四分位数评价结果见表 7。

4.4 基于贝叶斯统计分位数回归分析中的测试集的预测效果

贝叶斯统计分位数回归分析中的预测效果评价结果见表 8。

表5 基于经典统计分位数回归分析中的训练集所得到的四分位数评价结果

抽样次数	模型评价指标	*Abserror*	R^2	*MSE*
1	下四分位数(Q_1)	0.1564414064	0.07421	388061.74
	中位数(Q_2)	0.1423917874	0.08179	292838.70
	上四分位数(Q_3)	0.1792785241	0.05863	413232.82
2	下四分位数(Q_1)	0.1525651466	0.10717	366026.21
	中位数(Q_2)	0.1388277689	0.09827	284666.86
	上四分位数(Q_3)	0.1721253907	0.10063	388984.45
3	下四分位数(Q_1)	0.1463320482	0.09383	362512.09
	中位数(Q_2)	0.1321188585	0.08577	275975.13
	上四分位数(Q_3)	0.1660328832	0.07459	386335.27
4	下四分位数(Q_1)	0.1523820114	0.09686	371740.23
	中位数(Q_2)	0.1391723299	0.08534	284278.70
	上四分位数(Q_3)	0.1749863296	0.08107	397050.55
5	下四分位数(Q_1)	0.1524875811	0.08011	365643.79
	中位数(Q_2)	0.1384267716	0.08003	279577.61
	上四分位数(Q_3)	0.1735386080	0.04813	391024.08
6	下四分位数(Q_1)	0.1536339210	0.07520	382597.67
	中位数(Q_2)	0.1393904644	0.07071	289592.58
	上四分位数(Q_3)	0.1741722021	0.06464	401501.87
7	下四分位数(Q_1)	0.1517214804	0.12015	375126.87
	中位数(Q_2)	0.1379998017	0.08945	285011.54
	上四分位数(Q_3)	0.1744036945	0.07434	399188.06
8	下四分位数(Q_1)	0.1546086458	0.10244	383779.18
	中位数(Q_2)	0.140947790	0.08723	292239.83
	上四分位数(Q_3)	0.1787636357	0.07329	411646.88
9	下四分位数(Q_1)	0.1510109362	0.13725	359408.94
	中位数(Q_2)	0.1387711610	0.10155	286503.15
	上四分位数(Q_3)	0.1741560854	0.08734	399317.85
10	下四分位数(Q_1)	0.1555281103	0.07566	382720.42
	中位数(Q_2)	0.1413210001	0.07679	293239.13
	上四分位数(Q_3)	0.1757461463	0.08541	403685.18

表6 基于经典统计分位数回归分析中的预测效果评价结果

抽样次数	模型评价指标	*Abserror*	*NMSE*	*MSE*
1	下四分位数(Q_1)	0.1499364150	1.34025	459156.08
	中位数(Q_2)	0.1094981991	0.94176	260505.17
	上四分位数(Q_3)	0.1321630393	0.72267	315430.96
2	下四分位数(Q_1)	0.1382515050	0.90009	376628.53
	中位数(Q_2)	0.1448794009	1.04354	345736.74
	上四分位数(Q_3)	0.1919251339	1.18185	529392.41
3	下四分位数(Q_1)	0.1952015483	0.78637	438710.47
	中位数(Q_2)	0.2250631030	1.14509	502581.54
	上四分位数(Q_3)	0.2683186438	1.09735	673243.71
4	下四分位数(Q_1)	0.1604250406	0.93765	458202.13
	中位数(Q_2)	0.1374250176	0.91389	334227.46
	上四分位数(Q_3)	0.1702685060	0.92648	462716.86
5	下四分位数(Q_1)	0.1638658208	0.82731	539341.34
	中位数(Q_2)	0.1433003988	0.88681	385929.63
	上四分位数(Q_3)	0.1655359549	0.77803	459419.01
6	下四分位数(Q_1)	0.1519677879	0.78045	363556.22
	中位数(Q_2)	0.1420659786	0.85112	301454.85
	上四分位数(Q_3)	0.1868531873	0.92985	445289.92
7	下四分位数(Q_1)	0.1675247153	1.07637	424440.44
	中位数(Q_2)	0.1484353130	0.93881	326369.72
	上四分位数(Q_3)	0.1781714160	0.96911	445526.18
8	下四分位数(Q_1)	0.1520162160	1.16962	404619.07
	中位数(Q_2)	0.1196951734	0.92021	257967.31
	上四分位数(Q_3)	0.1461373001	0.97998	353579.46
9	下四分位数(Q_1)	0.1404679762	0.87735	354862.15
	中位数(Q_2)	0.1427851878	1.12484	316683.15
	上四分位数(Q_3)	0.1889873195	1.19272	489738.10
10	下四分位数(Q_1)	0.1366446944	0.82220	370403.20
	中位数(Q_2)	0.1194356437	0.87146	261367.83
	上四分位数(Q_3)	0.1493279710	0.78074	351686.13

表7 贝叶斯统计分位数回归分析中的训练集四分位数评价结果

抽样次数	模型评价指标	*Abserror*	R^2	*MSE*
1	下四分位数(Q_1)	0.1447636491	0.23777	319504.87
	中位数(Q_2)	0.1412749621	0.08981	290282.81
	上四分位数(Q_3)	0.1530329989	0.27123	319904.28
2	下四分位数(Q_1)	0.1420651702	0.23908	319504.87
	中位数(Q_2)	0.1383528270	0.10706	290282.81
	上四分位数(Q_3)	0.1503343725	0.27868	319904.28
3	下四分位数(Q_1)	0.1345386069	0.24895	311946.86
	中位数(Q_2)	0.1430669981	-0.06804	281891.61
	上四分位数(Q_3)	0.1560533223	0.13806	311976.42
4	下四分位数(Q_1)	0.1418817418	0.24041	299780.60
	中位数(Q_2)	0.1385497159	0.09086	321682.33
	上四分位数(Q_3)	0.1507774847	0.27904	359033.25
5	下四分位数(Q_1)	0.1409091320	0.23226	312654.94
	中位数(Q_2)	0.1374859863	0.08612	282562.62
	上四分位数(Q_3)	0.1489151506	0.25640	311511.63
6	下四分位数(Q_1)	0.1419308845	0.23011	305167.06
	中位数(Q_2)	0.1383689866	0.07974	277728.03
	上四分位数(Q_3)	0.1510181922	0.25761	305468.36
7	下四分位数(Q_1)	0.1408657931	0.26187	318510.83
	中位数(Q_2)	0.1372789905	0.09370	286778.46
	上四分位数(Q_3)	0.1501501560	0.26812	318672.07
8	下四分位数(Q_1)	0.1434524080	0.24980	314703.12
	中位数(Q_2)	0.1401782497	0.09370	283680.56
	上四分位数(Q_3)	0.1526909099	0.27949	315620.91
9	下四分位数(Q_1)	0.1421246198	0.24317	320770.43
	中位数(Q_2)	0.1381404935	0.10691	290167.18
	上四分位数(Q_3)	0.1502783122	0.27868	320055.13
10	下四分位数(Q_1)	0.1446333295	0.22107	315281.75
	中位数(Q_2)	0.1404584334	0.08462	284791.21
	上四分位数(Q_3)	0.1531065638	0.26944	315599.58

表8 贝叶斯统计分位数回归分析中的预测效果评价结果

抽样次数	模型评价指标	*Abserror*	*NMSE*	*MSE*
1	下四分位数(Q_1)	0.1207128874	0.98315	336818.44
	中位数(Q_2)	0.1077682459	0.94350	260986.15
	上四分位数(Q_3)	0.1127535475	0.57546	251176.93
2	下四分位数(Q_1)	0.1347942039	0.82820	346545.99
	中位数(Q_2)	0.1418622950	1.03787	343858.59
	上四分位数(Q_3)	0.1629903327	0.91785	411139.63
3	下四分位数(Q_1)	0.1950545946	0.72790	406086.52
	中位数(Q_2)	0.2203860697	1.11853	490922.94
	上四分位数(Q_3)	0.2376528133	0.87913	539361.38
4	下四分位数(Q_1)	0.1431932981	0.75485	368870.56
	中位数(Q_2)	0.1376158986	0.92158	337039.78
	上四分位数(Q_3)	0.1453264877	0.74045	369810.17
5	下四分位数(Q_1)	0.1500819081	0.68477	446413.95
	中位数(Q_2)	0.1426032992	0.88286	384212.98
	上四分位数(Q_3)	0.1463296107	0.64870	383052.48
6	下四分位数(Q_1)	0.1432671353	0.68453	318871.81
	中位数(Q_2)	0.1396234358	0.83683	296394.40
	上四分位数(Q_3)	0.1566924659	0.71717	343443.03
7	下四分位数(Q_1)	0.1530798767	0.90375	356370.54
	中位数(Q_2)	0.1468749457	0.93339	324486.18
	上四分位数(Q_3)	0.1581795318	0.77285	355299.85
8	下四分位数(Q_1)	0.1316991702	0.90174	311950.38
	中位数(Q_2)	0.1194627792	0.92493	259291.49
	上四分位数(Q_3)	0.1240821635	0.75904	273865.75
9	下四分位数(Q_1)	0.1387306503	0.81333	328967.22
	中位数(Q_2)	0.1416365251	1.13195	318683.10
	上四分位数(Q_3)	0.1613875381	0.92056	377987.05
10	下四分位数(Q_1)	0.1234630521	0.67042	302026.15
	中位数(Q_2)	0.1163881826	0.84681	253975.96
	上四分位数(Q_3)	0.1259272871	0.60494	272497.80

4.5 基于经典统计与贝叶斯统计分位数回归评价指标的比较

基于经典统计与贝叶斯统计的拟合效果评价指标均值比较见表 9。

表 9 基于经典统计与贝叶斯统计的拟合效果评价指标均值比较

评价指标	分位数	经典统计学	贝叶斯统计学
Abserror 均值	下四分位数(Q_1)	0.152671	0.141717
	中位数(Q_2)	0.138937	0.139316
	上四分位数(Q_3)	0.174320	0.151636
R^2 均值	下四分位数(Q_1)	0.096288	0.240449
	中位数(Q_2)	0.085693	0.076448
	上四分位数(Q_3)	0.074807	0.257675
MSE 均值	下四分位数(Q_1)	373761.7	313782.5
	中位数(Q_2)	286392.3	288984.8
	上四分位数(Q_3)	399196.7	319774.6

经典统计与贝叶斯统计预测效果评价指标均值比较见表 10。

表 10 经典统计与贝叶斯统计预测效果评价指标均值比较

评价指标	分位数	经典统计学	贝叶斯统计学
Abserror 均值	下四分位数(Q_1)	0.155630	0.143408
	中位数(Q_2)	0.143258	0.141422
	上四分位数(Q_3)	0.177769	0.153132
NMSE 均值	下四分位数(Q_1)	0.951766	0.795264
	中位数(Q_2)	0.963753	0.957825
	上四分位数(Q_3)	0.955878	0.753615
MSE 均值	下四分位数(Q_1)	418992.0	352292.2
	中位数(Q_2)	329282.3	326985.2
	上四分位数(Q_3)	452602.3	357763.4

5 讨论

表 9、10 分别表示在十折交叉验证中,训练集的拟合效果和测试集的预测效果。对于拟合效果来讲,贝叶斯统计的 Q_1 和 Q_3 的 *Abserror* 均值和 *MSE* 均值小于经典统计的指标,R^2 则大于经典统计,在 Q_2 的指标上恰恰相反。这说明在拟合效果上,贝叶斯统计在下、上四分位数的拟合效果要优于经典统计,而对中位数的拟合效果经典统计要优于贝叶斯统计。

对于预测效果来讲,基于贝叶斯统计的各个分位数的各项反映误差的指标取值均小于经典统计的相应指标的取值,说明贝叶斯统计对于分位数回归分析的预测效果要优于经典统计的预测效果。

因此,我们认为在分位数回归分析中,从评价指标数值上来讲,基于贝叶斯分位数回归分析的 MCMC 过程要比经典统计中的 QUANTREG 过程拟合效果和预测效果更好。但贝叶斯分位数回归分析也有其局限性:①样本量过大时,计算量庞大,耗时久,尤其在十折交叉验证过程中,更加费时费力;②贝叶斯统计的分位数回归分析方法尚不完善,对于自变量的筛选暂无好的方法;③贝叶斯统计的分位数回归分析的预测值无法自动计算,需要手动输入计算。

【参考文献】

[1] 何凤霞,王凤竹. 分位数回归及其在 R 中的实现[J]. 湖南文理学院学报(自然科学版),2013,25(3):10-15.

[2] Koenker R, Bassett G. Regression quantiles[J]. Econometrica, 1978, 46(1): 33-50.

[3] 曾惠芳,朱慧明. 基于 MCMC 算法的贝叶斯分位回归计量模型及应用研究[D]. 长沙:湖南大学,2011.

[4] 胡良平. 医学统计学——运用三型理论进行现代回归分析[M]. 北京:人民军医出版社,2010:98.

[5] 茆诗松. 贝叶斯统计[M]. 2 版. 北京:中国统计出版社,1999:3-4.

[6] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 3 版. 北京:中国人民大学出版社,2012:42.

[7] 胡良平,胡纯严. SAS 语言基础与高级编程技术[M]. 北京:电子工业出版社,2014:349.

(孙承媛 编辑 2018-02-05 收稿)