Statistics and Application 统计学与应用, 2018, 7(6), 605-613
Published Online December 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.76069

Hans汉斯

Study on the Health Insurance Premiums Income Based on Quartile Autoregression Model

Shifeng Gong, Yue Wang, Haomin Zhang*

Guilin University of Technology, Guilin Guangxi
Email: gongshifeng233@163.com, 875848690@qq.com, *zhanghm@glut.edu.cn

Received: Nov. 20th, 2018; accepted: Dec. 5th, 2018; published: Dec. 12th, 2018

Abstract

Health insurance premium income is a key economic indicator to measure the development of health insurance industry. In this paper, the time series data of health insurance premium income from January 1991 to June 2017 were modeled with the quantile regression method. Firstly, the model AR(3) was identified, and then the quantile autoregressive model was established by the method of quantile regression, and the increase and decrease trend of the original data was fitted accurately. Finally, the short-term prediction is made by autoregressive AR(3) model and quantile autoregressive QAR(3) model respectively. The prediction results of the two models were compared based on several evaluation indicators. The results show that the quantile autoregressive model is more effective.

Keywords

Health Insurance Premium Income, Time Series, Quantile Autoregressive, Forecast

基于分位数自回归模型的健康险保费收入研究

龚石凤，王　越，张浩敏*

桂林理工大学，广西　桂林
Email: gongshifeng233@163.com, 875848690@qq.com, *zhanghm@glut.edu.cn

收稿日期：2018年11月20日；录用日期：2018年12月5日；发布日期：2018年12月12日

摘　要

健康险保费收入是度量健康保险业发展情况的关键经济指标。本文结合分位数回归方法对我国1991年1月~2017年6月健康险保费收入的时序数据建模：首先识别了AR(3)模型，之后运用分位数回归方法建立分位

*通讯作者。

文章引用：龚石凤，王越，张浩敏．基于分位数自回归模型的健康险保费收入研究[J]．统计学与应用，2018，7(6): 605-613. DOI: 10.12677/sa.2018.76069

数自回归模型，并较准确地拟合出原数据的增减趋势，最后用自回归AR(3)模型和分位数自回归QAR(3)模型分别做短期预测，基于多个评价指标比较两种模型预测效果，结果表明分位数自回归模型预测效果更好。

关键词

健康险保费收入，时间序列，分位数回归，预测

1. 引言

我国的健康保险业务自上世纪 80 年代恢复以来，得到了快速的发展，根据中国保监会发布的健康保险保费收入数据，我国健康保险保收入由 1999 年的 3.65 亿多元增加到 2017 年的 4389.4 亿多元，期间平均每年增长约为 34.16%。在新时代背景下，健康与养老服务业已逐渐成为我国新的经济增长点和扩大内需的着力点，我国健康保险的发展也面临着更多新的挑战。国家和保险业需要根据健康保险的现状及时做出相应的措施和调整战略。

Koenker 和 Bassett (1978) [1]提出的分位数回归是最小二乘回归方法的优化[2]。与普通的均值回归相比，分位数回归能精细地刻画自变量对于响应变量对应于不同分位点的不同影响。将分位数回归模型与时间序列分析结合能较好地提高模型的预测能力和实用性[1] [2] [3] [4]。彭良玉等(2011)对分位数回归和时间序列理论进行了深入研究，并在此基础上分析了澳大利亚月度红酒销量数据，认为与时间序列模型相比，分位数回归方法能够得到更加完整的红酒销量信息[3]。盛选义等(2012)将分位数回归方法应用到时间序列系数求解中，分析我国对外贸易总额数据，实例验证结果表明模型预测效果较好且具有一定的应变能力[4]。崔丙维(2013)根据时间序列的一般理论识别了 AR 模型，然后按照将所得模型结合分位数回归的思路，建立了分位数自回归模型，还将此模型应用到考察风速变化的实际问题中，结果显示用分位数回归模型进行拟合能有更高的拟合度[5]。本文拟运用分位数自回归(QAR)模型研究健康险保费收入的预测问题。

2. 实证分析

本文对我国健康保险保费收入实际数据进行分析，为了更全面且准确地探究我国健康险保费收入的情况，从国泰安数据库[6]选用了 1999 年 1 月到 2018 年 2 月我国的健康险保费收入数据，这是官方发布的有记录以来的所有数据。将原始数据分为两个部分，1999 年 1 月直至 2017 年 6 月的保费收入数据组成第一部分，作为建立模型的依据，剩余的 2017 年 7 月至 2018 年 2 月的健康险保费收入共 8 个数据用来检验模型预测效果。

2.1. 数据预处理

图 1 为我国健康险保费收入时间序列图。可以看出该序列值基本是逐年增加但近年的最高值与最低值差距较大。图 2 为健康保险保费收入时间序列的分解图。可以看出序列不仅具有逐年增加的趋向，而且所显示出来的季节性也不可忽视。从而可以判断该序列是非平稳的，因此在建模之前要进行必要的预处理以使其稳定。

由于选用的是月度数据，尝试对原序列进行 1 阶 12 步差分，差分后序列显示均值平稳但在序列后面部分方差较大，如图 3 所示，这是序列存在异方差才会表现的特点。为了进一步确认序列是否真的存在异方差性，考察 1 阶 12 步差分后序列残差平方图的方法，所得结果如图 4 所示，可以看到图中曲线后部的差异更加显而易见，因此得出残差序列存在异方差的结论。

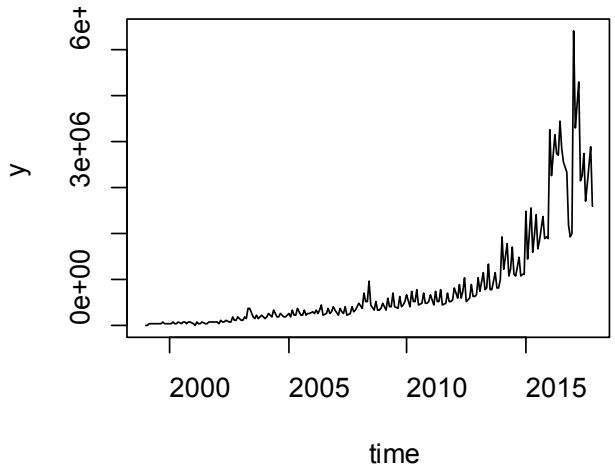

Figure 1. Time series chart of China's health insurance premium income
图 1. 我国健康险保费收入时序图

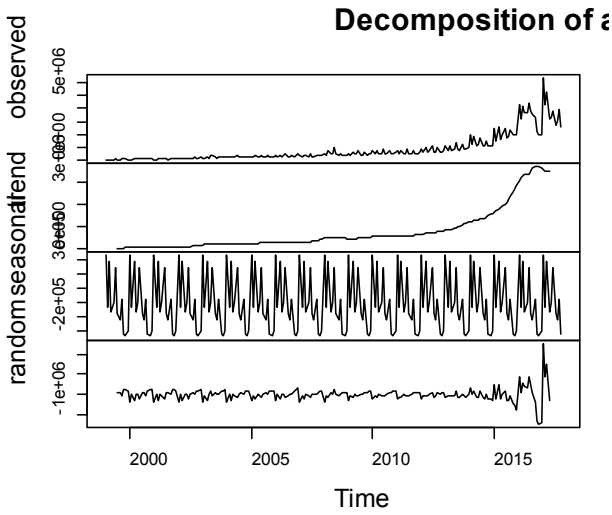

Figure 2. Data decomposition map of China's health insurance premium income
图 2. 我国健康保险保费收入数据分解图

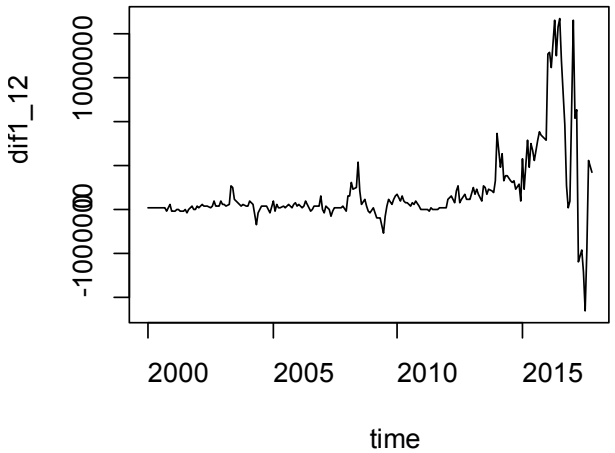

Figure 3. Time series diagram of the 1st order 12-step difference graph
图 3. 1 阶 12 步差分后时序图

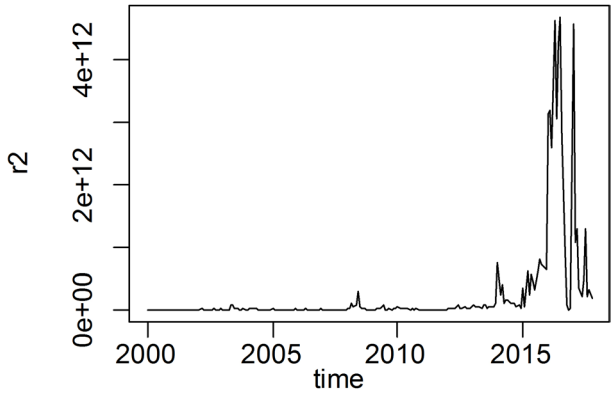

Figure 4. 1st order 12-step difference residual squared graph
图 4. 1 阶 12 步差分后残差平方图

下面通过对数变换来消除序列的异方差性。图 5 是经过对数变换的序列 1 阶 12 步差分后的时间序列图。图中曲线大致上是在一个常数值附近进行无规律的变动，并且波动没有离开该常数很远。按照时间序列图检验法的规则可以认为处理后的序列已经平稳。但仅根据时序图判断序列平稳性存在一定的主观因素，为了得到更加客观的结果，接下来使用单位根法检验处理后的时间序列的平稳性，该检验得到的 p 值是 0.04354，显然在显著性水平取 α 的值是 0.05 时拒绝原假设，所以单位根检验法得到的结果同样证实了差分后序列是平稳的。

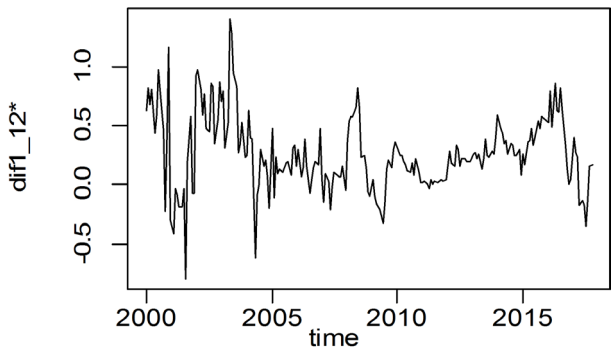

Figure 5. Time-series sequence of 1st order 12-step difference after logarithmic transformation
图 5. 经过对数变换的序列 1 阶 12 步差分后时序图

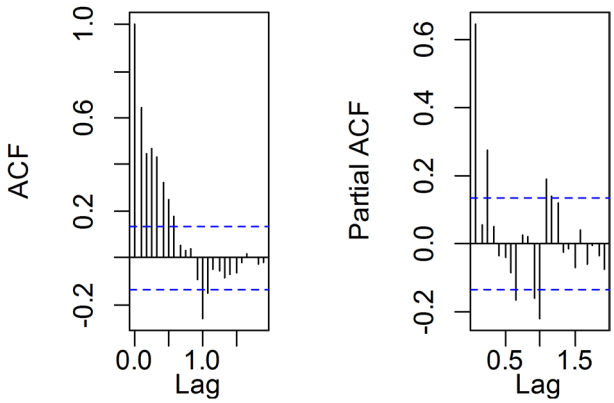

Figure 6. Autocorrelation coefficient and partial autocorrelation coefficient of pre-processed sequence
图 6. 预处理后序列的自相关系数及偏自相关系数图

2.2. 平稳序列的白噪声检验

经过预处理之后所得序列已经平稳，下一步对平稳序列进行纯随机性检验。采用 Q_{LB} 检验统计量做平稳序列的白噪声检验，结果见表 1。

Table 1. Parameter table of white noise test
表 1. 白噪声检验参数表

延迟阶数	Q_{LB} 统计量检验	
	Q_{LB} 统计量的值	p 值
6	243.8	<0.0001
12	264.97	<0.0001

如表 1 中的结果显示，延迟 6 期延迟 12 期的 Q_{LB} 统计量对应的 p 值都远小于显著性水平，可以判定差分后序列为非白噪声序列，即该序列中隐含着有用信息，有必要继续对此序列做进一步分析建模。

2.3. 模型识别

按照前文预处理后得到的平稳序列，其样本自相关及偏自相关图如图 6 所示，显然图 6 中显示的样本自相关图具有拖尾的特点，而偏自相关图则是三阶截尾，根据模型定阶准则初步认为选择 AR(3)模型是最合理的。但是这里可以选择的模型并不只有 AR(3)一个，还存在其他有效的模型，本文根据 AIC 准则判断最优模型，计算各个有效模型的 AIC 函数值，所得结果如表 2 所示。

Table 2. AIC values for each valid model
表 2. 各个有效模型的 AIC 值

模型	AR(1)	AR(2)	AR(3)	ARMA(1,1)	MA(1)	MA(2)	MA(3)
AIC	15.68	16.14	1.2	12.7	43.74	32.54	30.44

依据 AIC 函数值最小对应的模型即为最优的判断标准，比较表 2 中每一个模型的 AIC 函数值可以确认选用 AR(3)模型是最恰当的，与按照 ARMA 模型定阶原则得到的模型一致。AR(3)模型形式如式(2-1)所示。

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \varepsilon_t . \tag{2-1}$$

由于后面会比较时间序列的自回归模型和时间序列分位数回归模型的预测效果，所以这里用最小二乘估计方法，计算了 AR(3)模型的参数，结果如式(2-2)。

$$y_t = 0.06052 + 0.57761 y_{t-1} - 0.1085 y_{t-2} + 0.28558 y_{t-3} + \varepsilon_t . \tag{2-2}$$

Table 3. Results of the t-test of the parameters
表 3. 参数 t 检验结果

系数	β_0	β_1	β_2	β_3
p 值	0.0063	<0.001	0.01353	<0.001

表 3 给出了对 AR(3)模型系数的 t 检验结果，可以看到所有系数的 p 值都小于显著性水平 0.05，也就是说当期的健康保险保费收入受其前三期的健康险保费收入影响，并且影响是显著的。

2.4. 分位数回归

将分位数回归运用到自回归 AR(3)模型上来，首先要选择合适的分位数，在实际应用中，中位数($\tau=0.5$)有着非常重要的作用，常常和均值共同反映样本数据所包含的位置信息。此外，中位数的取值是由其在所有标志值中所处位置所决定的，不受分布数列极值的影响，相对于均值来说，中位数在一定程度上提高了对样本分布数列的代表性；当少数误差数据严重偏离真实数据时，会导致样本均值大受影响，而丝毫不会影响中位数[7]。所以本文选取 5 个分位数 $\tau=0.05,0.15,0.5,0.85,0.95$，其中 $\tau=0.5$ 的分位线即为中位数回归，为后文的点预测(主要指中位数回归预测)和区间预测($\tau=(0.05,0.95)$ 表示 $\tau=0.05$ 预测区间，$\tau=(0.15,0.85)$ 表示 $\tau=0.15$ 预测区间)奠定基础。

估计不同分位点条件下模型的系数 $\beta_0,\beta_1,\beta_2,\beta_3$，得到系数的估计值如表 4 所示。观察表 4 可知不同的分位点所对应的系数不同，得到了时间序列分位数回归模型(QAR)。

Table 4. Estimates of coefficients of models corresponding to different quantiles
表 4. 不同分位点对应的 AR(3)系数估计值

分位点	β_0	β_1	β_2	β_3
$\tau=0.05$	−0.18666	0.42399	−0.14945	0.19389
$\tau=0.15$	−0.08361	0.59103	−0.11712	0.15030
$\tau=0.5$	0.03036	0.70617	−0.03864	0.18706
$\tau=0.85$	0.1989	0.78035	−0.23647	0.39172
$\tau=0.95$	0.47109	0.61673	−0.26268	0.30336

以中位数回归模型为例，健康险保费收入在 t 时刻拟合的分位数自回归模型(QAR)为：

$$Q_{0.5}\left(y_t \mid F_{t-1}\right)=0.03036+0.70617y_{t-1}-0.03864y_{t-2}+0.18706y_{t-3}. \tag{2-3}$$

其中 F_{t-1} 表示 y_t 的滞后项产生的 σ-域，这里 $F_{t-1}=\{y_{t-1},y_{t-2},y_{t-3}\}$。

对分位点与模型的系数进行研究。图 7 描述的是模型各系数随分位数变化而产生相应变化的情况，即对 AR(3)模型实施分位数回归分析后得到的结果，图 7 中的黑色虚线是在不同的分位点处各参数的回归系数，灰色形成的阴影范围为估计值的 95%置信区间带[1]。而红色的三条平行线中，实线是最小二乘法所得的参数估计值而虚线则是其 95%的置信区间上下限。

根据自回归模型 $y_t=\beta_0+\beta_1y_{t-1}+\beta_2y_{t-2}+\beta_3y_{t-3}+\varepsilon_t$ 可知，图 7 中第一个图的纵坐标为常数项 β_0 的估计值，第二个图的纵坐标为 y_{t-1} 的系数 β_1 的估计值，第三个图的纵坐标为 y_{t-2} 的系数 β_2 的估计值，第四个图的纵坐标则是 y_{t-3} 的系数 β_3 的估计值。分位数刚开始上升时，参数 β_1 的估计值先存在小段下降的情况而后逐渐上升。也就是说，当健康保险发展低迷时，对于经过对数转换和差分后的数据，前一期对后一期的影响会随这分位数的增加而减少，而在正常发展水平下，前后两期数据之间的影响随着分位数的增加而增加。而参数 β_2 随着分位数的增加其先增加然后持续平稳而后下降，说明只有在健康保险发展很不好时，分位数增加 y_{t-2} 对 y_t 的影响是正向的，或者健康保险收入走高，y_{t-2} 对 y_t 的影响则是分位数增加而下降。再看参数 β_3，其与分位数之间的关系则较为简单，分位数逐渐增加 β_3 大体也是上升的趋势，即 y_{t-3} 对 y_t 的影响会随着分位数的增加而变大。

根据已得出参数估计值的时间序列分位数回归模型，作不同分位点条件下的模型拟合图如图 8 所示。图中上下两条红色线分别代表 0.95 和 0.05 分位点条件下的拟合值，绿色线由上往下分别代表 0.85 和 0.15 分位点条件下的拟合值，蓝色线则代表 0.5 分位点条件下的拟合值，黑色线为经过预处理的数据。从整

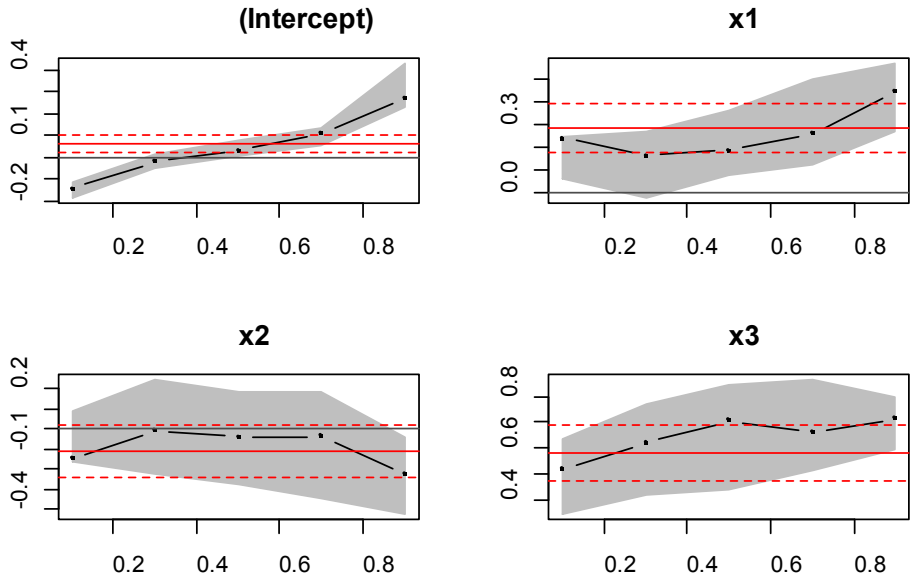

Figure 7. Trends and confidence intervals for model parameters
图 7. 模型参数的变化趋势及置信区间

体图可以看出，基于时间序列的分位数回归模型的拟合结果均能合理的反映出健康险保费收入的变化规律，此外中位数($\tau = 0.5$)的分位数曲线与真实健康险保费收入的值最为接近，这与分位数回归的概念相符。

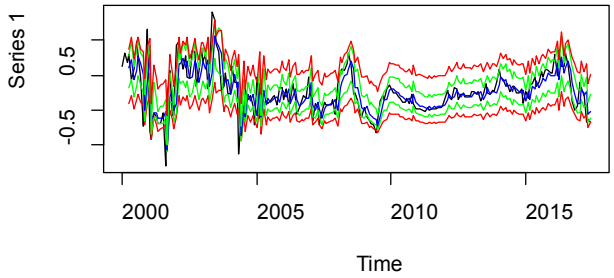

Figure 8. Quantile autoregressive model fitting effect diagram
图 8. 分位数自回归模型拟合效果图

2.5. 序列预测

利用最小二乘法和分位数回归估计得到的模型预测 2017 年 7 月到 2018 年 2 月的健康险保费收入。为了更好比较预测效果，表中给出了 2017 年 7 月到 2018 年 2 月的实际观测值，得到的健康险保费收入预测值如表 5 所示。

Table 5. Predicted value of health insurance premium income (ten thousand yuan)
表 5. 健康险保费收入预测值(万元)

时间	实际值	OLS	$\tau = 0.05$	$\tau = 0.15$	$\tau = 0.5$	$\tau = 0.85$	$\tau = 0.95$
2017.7	2,707,473	3,575,774	3,176,085	3,433,368	3,710,793	4,296,925	5,909,880
2017.8	3,084,893	3,573,493	2,347,428	2,815,124	3,458,874	5,304,820	11,791,011
2017.9	3,881,245	3,643,690	1,726,717	2,296,265	3,301,410	7,087,132	25,060,086
2017.10	2,590,383	3,834,533	1,248,183	1,861,270	3,230,513	10,433,442	59,240,017

Continued

2017.11	2,634,552	4,186,069	891,104.9	1,503,113	3,242,421	17,457,162	163,165,118
2017.12	2,839,169	4,678,434	633,895.1	1,211,176	3,324,774	32,535,723	221,905,174
2018.1	5,323,376	5,326,937	449,555.7	974,154.8	3,473,874	66,476,508	259,300,350
2018.2	3,426,597	6,179,651	317,814.8	782,439.9	3,692,591	149,564,012	718,467,278

判断模型点预测效果的模型评价指标有[7]：

1) 根均方误差 $\mathrm{RMSE}(\tau)$

$$\mathrm{RMSE}(\tau)=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t-\hat{Q}_\tau\left(y_t\mid F_{t-1}\right)\right)^2}\ . \tag{2-4}$$

2) 平均绝对误差 $\mathrm{MAE}(\tau)$

$$\mathrm{MAE}(\tau)=\frac{1}{T}\sum_{t=1}^{T}\left|y_t-\hat{Q}_\tau\left(y_t\mid F_{t-1}\right)\right|. \tag{2-5}$$

3) 平均百分比误差 $\mathrm{MAPE}(\tau)$

$$\mathrm{MAPE}(\tau)=\frac{1}{T}\sum_{t=1}^{T}\left|\frac{y_t-\hat{Q}_\tau\left(y_t\mid F_{t-1}\right)}{y_t}\right|\times 100\%\,. \tag{2-6}$$

其中，T 是预测值的个数，Y_t 表示实际观测的数据，$\hat{Y}_t$ 表示用模型所得的预测值，F_{t-1} 表示 y_t 的各阶滞后变量生成的 σ-域。表 6 列出了自回归模型(AR)、分位数自回归模型(QAR)对健康险保费收入数据点预测($\tau=0.5$)下的预测值的效果比较情况。显然 QAR 模型相对于 AR 模型在预测效果上有了显著提高，QAR 模型的 RMSE 提升了 38.7%、MAE 提升了 35.8%，MAPE 更为显著提升了 43.9%，同时可以说明简单地运用均值回归模型对数据进行预测缺乏合理有效性。

Table 6. Comparison of point prediction results
表 6. 点预测效果比较

模型	RMSE	效果提升百分比	MAE	效果提升百分比	MAPE	效果提升百分比
AR(p)	1,412,729.633	-	1,123,250.344	-	38.26783	-
QAR(p)	865,246.7636	38.7%	725,779.5013	35.8%	21.43925	43.9%

进一步考察区间预测的效果，评价指标有 $\mathrm{FICP}(\tau)$ [7]，它是指实际值落在预测区间的概率。

$$\mathrm{FICP}(\tau)=\frac{1}{T}\sum_{t=1}^{T}I_t(\tau)\times 100\%,\ \tau\in(0,0.5)\,. \tag{2-7}$$

其中 $I_t(\tau)$ 为 Boolean 变量，当 Y_t 落入其预测区间 $\left[\hat{Q}_\tau\left(y_t\mid F_{t-1}\right),\hat{Q}_{1-\tau}\left(y_t\mid F_{t-1}\right)\right]$ 时，$I_t(\tau)=1$，否则 $I_t(\tau)=0$。根据表 5，2017 年 7 月至 2018 年 2 月的 8 个实际值，仅有 2017 年 7 月的真实值未落入 $\tau=0.05$ 及 $\tau=0.15$ 预测区间，$\tau=0.05$ 及 $\tau=0.15$ 预测区间内的覆盖率达到了 87.5%，说明分位数自回归模型区间预测的效果较好。

3. 结论

本文按照通常的时间序列分析方法对数据建模之后，在选定分位数的条件下获得时间序列分位数回归模型，并以此模型进行参数估计和预测，主要将这个方法应用于我国健康险保费收入的研究中。采用

1999 年 1 月到 2017 年 6 月我国健康保险保费收入进行建模，识别得到 AR(3)模型，从选定的模型看，我国健康险当月的保费收入会受此前三个月保费收入的影响。一方面可以理解为健康险保费收入存在延展性，这是因为长期健康保险保单一般会有分期缴纳的保费，所以保费收入会表现出一定的滞后效应；另一方面保险公司可以优化健康保险的市场格局，升级产品结构，从而增强自身竞争力寻求发展。

根据时间序列分位数回归模型进行拟合，拟合结果能合理地反映出健康险保费收入的变化规律，此外中位数($\tau = 0.5$)的分位数曲线与实际健康险保费收入的值最为接近，这与分位数回归的概念相符，也说明用 $\tau = 0.5$ 的模型进行点预测的合理性。用 AR(3)模型和时间序列分位数回归模型分别预测了 2017 年 7 月到 2018 年 2 月的健康险保费收入，分位数自回归(QAR)模型的点预测效果相对于自回归(AR)模型有了显著提高，QAR 模型的根均方误差 RMSE 提升了 38.7%、平均绝对误差 MAE 提升了 35.8%，平均百分比误差 MAPE 更为显著提升了 43.9%，显然分位数条件下的自回归模型的点预测效果比用时间序列模型所得的效果要好得多。最后还考察了区间预测的效果，$\tau = 0.05$ 及 $\tau = 0.15$ 预测区间内的覆盖率达到了 87.5%，表明分位数自回归模型区间预测的效果较好。经过对健康险保费收入的实证分析，表明分位数回归法不但能够有效地估计模型参数，并且能够得到更好的预测效果。

基金项目

国家自然科学基金项目资助(No.61763008, 71762008)；广西自然科学基金项目资助(No.2106GXNSFAA380194)。

参考文献

[1] Roger, K. and Bassett, G. (1978) Regression Quantiles. *Econometrica*, **64**.

[2] 朱平芳, 张征宇. 无条件分位数回归: 文献综述与应用实例[J]. 统计研究, 2012, 29(3): 88-96.

[3] 彭良玉. 分位数回归在时间序列中的应用[D]: [硕士学位论文]. 天津: 天津大学, 2010.

[4] 盛选义, 彭良玉. 分位数回归在时间序列中的应用[J]. 太原师范学院学报(自然科学版), 2011, 10(3): 25-29.

[5] 崔丙维. 基于分位数回归的时间序列模型及应用[D]: [硕士学位论文]. 北京: 华北电力大学, 2013.

[6] 国泰安 SCMAR 数据库. 数据中心[EB/OL]. http://www.gtarsc.com/Home

[7] 杜艳芳. 基于分位数回归的空气质量指数分析[D]: [硕士学位论文]. 兰州: 兰州大学, 2016.