基于分位数回归的期刊影响因子影响因素研究[1]

[1]俞立平 [2]潘云涛 [2]武夷山

[1]宁波大学商学院 宁波 315211

[2]中国科学技术信息研究所 北京 100038

摘要：为了详细分析学术期刊影响因子在不同水平下的影响因素及其特点，本文利用分位数回归模型，分析了平均引文数、平均作者数、地区分布数、海外论文比、基金论文比、期刊时效性对影响因子的影响，结果表明，海外论文比与影响因子无关；高影响因子和低影响因子期刊，平均引文数与期刊影响因子无关；高影响因子期刊，基金论文比与影响因子无关。影响因子影响因素分析有助于期刊评价指标选取。对于不同水平期刊的评价，评价方法可能不同。

关键词：学术期刊 影响因子 分位数回归 影响因素

中图分类号：G304

Study on influence to academic journal impact factor based on Quantile Regression

Yu Liping[1], Pan Yuntao [2], Wu Yishan[2]

[1] Business school, Ningbo University, Ningbo (P.R. China)

[2] Institute of Scientific & Technical Information of China, Beijing (P.R. China)

Abstract: This paper analyzes influence of average average references per article，average number of authors，numbers of provinces，share of overseas contributions in total papers，share of grant-supported papers in total papers to impact factor based on quantile regression in order to assay impact factor's influence at different level. The results show that share of overseas contributions in total papers is independent of impact factor. Average references per article is independent of impact factor when impact factor is higher or lower. Analyzing of impact factor's influence is conducive to academic journal evaluation indicators. Evaluation of different levels should choose different evaluation methods.

Keywords: academic journal; impact factor; quantile regression; inflence

1 引言

美国著名情报学家加菲尔德博士在 20 世纪 60 年代对期刊文献的引文进行了大规模统计分析，得到了大量被引用文献集中在少数期刊上，而少量被引用文献散布在大量期刊中的结论，这是国外期刊评价理论的起源。国内外学者在期刊评价中设计了许多评价指标，典型的有 RCR(Schubert et al.,1983)[1]、RI、RW、PI、PW (Peter Vinkler,1986)[2]、NMCR（Braun & Glanzel,1990）[3]、FCS_m（Moed et al.,1995）[4]、H 指数（Hirsch,2005）[5]等。

[1]国家自然科学基金资助：发达国家科技期刊建设同经济实力、科技发展的关系暨期刊语言选择的历时性研究及借鉴意义（70973118）；ISTIC-THOMSON 科学计量学联合实验室开放基金项目：中国科研机构评价研究（IT2009001）。

作者简介：俞立平（1967-）男，江苏泰州人，博士，中国科学技术信息研究所博士后，宁波大学商学院教授，主要从事信息经济、科学计量领域的教学科研工作。发表论文 100 余篇，Email: chinayangzhou@yahoo.com.cn

潘云涛（1967-）女，研究员，主要从事科技评价领域的研究。发表论文近 70 余篇。

武夷山（1958-）男，研究员，中国科学技术信息研究所总工程师，主要从事情报学、科技管理、科学计量学领域的研究。发表论文、专著、译著等 500 余篇（部）。

学术期刊评价指标可以进一步分为期刊影响指标与期刊来源指标，所谓影响指标，就是指与期刊被引相关的指标，如总被引频次、影响因子、即年指标、学科影响指标、学科扩散因子、H 指数等等；期刊来源指标，就是指与期刊自身特性相关的一些指标，如基金论文比、平均引文数、平均作者数、地区分布数等等。还有一些指标兼顾影响指标和来源指标的特点，但更多地反映了期刊的时效性，如即年指标、引用半衰期、被引半衰期等。

在所有学术期刊评价指标中，影响因子作为期刊影响力的指标具有十分重要的地位，也可以说，建立在影响因子基础上的期刊和论文评价是文献计量学的基石。一般认为，影响因子越高的期刊，其学术质量也越高。影响因子和学术期刊来源指标、时效性指标之间存在着相关关系，对期刊影响因子的影响因素进行分析，有利于弄清它们的内在规律，从而对学术期刊评价工作进行指导。

在学术期刊评价指标内在关系研究领域，于挨福、马虎兆（2008）[6]利用面板数据分析了基金论文比、论文篇幅长短、期刊主办单位的实力和影响力、期刊的类型以及作者学术水平等因素对影响因子的影响。姜春林（2007）[7]研究了期刊 H 指数与影响因子的关系。刘雪立、董建军等（2006）[8]研究了中国医学期刊出版周期与即年指标关系。金碧辉、汪寿阳等（1999）[9]讨论了期刊学术质量与影响因子之间的关系，认为期刊影响因子与论文学术质量有直接联系，同时必须结合同行评议进行分析。迄今为止，总体上对期刊影响因子影响因素的实证研究不多，此外采用回归分析原理的研究较少，采用分位数回归的研究则未见报道。

普通回归本质上是一种均值回归，它能提供期刊影响因子影响因素的均值分析结果，但是，它又是一种相对粗糙的研究方法，分位数回归能够提供期刊不同水平影响因子影响因素的分析结果，处理方法更为精确。本文利用中国科学技术信息研究所的医学期刊数据，系统进行学术期刊影响因子影响因素的分析。

2研究方法

2.1 研究基本假设

假设一：基金论文比与期刊影响因子正相关。一般认为，能够通过激烈竞争获得科研基金的团队或个人，应该具有较强的实力。同时，有基金依托的论文，其研究条件相对较好，更容易出高质量的成果。因此，期刊基金论文比越高，期刊影响因子也应越高。

假设二：平均引文数与期刊影响因子正相关。论文的引文数越多，即参考文献越丰富，说明作者对前人的研究了解越多，“站在了巨人的肩膀上”，则这些论文质量应相对较好，期刊的影响因子越高。

假设三：平均作者数与期刊影响因子正相关。作者数较多的论文，能够集中大家的智慧，论文的质量一般也越高，其影响力越大。

假设四：地区分布数与影响因子正相关。期刊作者国内地区分布（主要指省级行政区）越多，说明期刊的覆盖面越广，其影响力也越大，也说明该刊对任何地区的作者都有吸引力，这是好期刊的特征，因此期刊影响因子和国内地区分布数呈正相关关系。

假设五：海外论文比与期刊影响因子正相关。能够吸引较多海外学者投稿的期刊，说明期刊质量较高，其影响力已经扩散到全球，因此，海外论文比越高，期刊影响因子越高。

假设六：期刊的时效性与期刊影响因子正相关。这里时效性并不是指期刊从投稿到刊出的发表周期，而是指期刊刊载论文的时效性。时效性越好的期刊，说明其论文较新，能跟踪学科前沿，而且，时效性较好的期刊对作者的吸引力很大，因此，时效性好的期刊的影响因子也越大。

当然，期刊影响因子的影响因素很多，由于数据的获得有一定的难度，暂且基于以上假设进行研究。

2.2 分位数回归模型

分位数回归是一种基于被解释变量 Y 的条件分布来拟合解释变量 X 的回归模型，是在均值回归上的拓展，最早由 Koenker、Basset（1978）[10]提出。它依据因变量的条件分位数对自变量 X 进行回归，这样得到了所有分位数下的回归模型。与普通最小二乘回归相比，分位数回归更能精确地描述自变量 X 对于因变量 Y 的变化范围以及条件分布形状的影响。

Koenker、Hallock（2001）[11]和Bernd、Peter（2007）[12]的研究认为，从理论上说，经典线性回归是拟合被解释变量Y的条件均值与解释变量X之间的线性关系，而分位数回归是通过估计被解释变量取不同分位数时，对特定分布的数据进行估计。最小二乘法估计的是解释变量对被解释变量的平均边际效果，而分位数回归估计的则是解释变量对被解释变量的某个特定分位数的边际效果。最小二乘法只能提供一个平均数，而分位数回归却能提供许多不同分位数的估计结果。

对于随机变量 Y 的一个随机样本$\{y_1,y_2,y_3,\cdots,y_n\}$，它的中位数线性回归就是求解使下面的绝对值偏差和为最小值：

$$\min_{\zeta} \sum |y_i - \zeta| \qquad (1)$$

中位数线性回归其实就是分位数线性回归的一个特例（τ=0.5），它在分位数线性回归中占有相当重要的地位，τ 分位数的样本分位数线性回归则是求满足 $\min_{\beta \in R^k} \sum_{-i} \rho\tau(y_i - x_i^{'}\beta(\tau))$ 的解 $\beta(\tau)$，它的展开式为：

$$\min_{\beta(\tau)\in R^k} \left[\sum_{(i:y_i \geq x_i^{'}\beta(\tau))} \tau |y_i - x_i^{'}\beta(\tau)| + \sum_{(i:y_i < x_i^{'}\beta(\tau))} (1-\tau)|y_i - x_i^{'}\beta(\tau)| \right] \quad \tau \in (0,1) \qquad (2)$$

在线性条件下，给定 x 后，Y 的 τ 分位数函数为：

$$Q_y(\tau | x) = x'\beta(\tau) \qquad \tau \in (0,1) \qquad (3)$$

在不同的 τ 下，就能得到不同的分位数函数。随着 τ 的值由 0 至 1，就能得到所有 y 在 x 上的条件分布轨迹，即一簇曲线，而不是像线性回归只能得到一条曲线。采用 Eviews6.0 软件可以方便地进行分位数回归估计。

3 数据

本文数据来自于中国科学技术信息研究所 CSTPC 数据库，中国科学技术信息研究所从 1987 年开始对中国科技人员在国内外发表论文数量和被引情况进行统计分析，并利用统计数据建立了中国科技论文与引文数据库，同时出版《中国科技期刊引证报告》。在分析期刊来源指标与影响因子关系时，由于不同学科期刊之间不可比，而在数据量较少的情况下进行研究又不具有代表性，因此本文选取期刊种类相对较多的医学期刊数据进行分析。数据是来自于 2008 年版《中国科技期刊引证报告》，数据是 2007 年的，共 537 种医学期刊。

期刊影响力指标有总被引频次，影响因子、扩散因子等，以影响因子最具有代表性，因此选取影响因子（YZ）作为期刊影响力指标。

表 1 数据描述统计量

变量	含义	均值	极大值	极小值	标准差
YZ	影响因子	0.43	1.71	0.01	0.26
JL	即年指标	0.05	0.31	0.00	0.04
BY	被引半衰期	7.63	11.40	0.88	1.13
YY	引用半衰期	4.31	7.24	0.29	0.84
YW	平均引文数	9.63	37.76	2.87	4.23

ZZ	平均作者数	4.02	6.48	2.00	0.83
DQ	地区分布数	23.38	31.00	2.00	5.91
HW	海外论文比	0．01	0.75	0.00	0.05
JJ	基金论文比	0.25	0.99	0.01	0.19
n	期刊数量		537		

期刊的时效性指标由 3 个指标构成：即年指标（JL）、被引半衰期（BY）、引用半衰期（YY），如果将 3 个变量放在同一方程回归，由于指标间的相关性较高，容易导致多重共线性，因此，先将 3 个指标进行标准化处理，每项指标最大值设为 100，其他数据按比例进行调整，得到标准化后的即年指标（JL’）、被引半衰期（BY’）、引用半衰期（YY’），考虑到被引半衰期和引用半衰期性质相似，3 个指标又各有特点，因此采取等权重原则，得到期刊的时效性指数 SX。即：

$$SX = \frac{1}{3}\mathrm{JL'} + \frac{1}{3}\mathrm{BY'} + \frac{1}{3}\mathrm{YY'} \qquad (4)$$

需要说明的是，被引半衰期和引用半衰期是两个反向指标，必须进行正向处理，方法是用 100 减去其标准化后的结果再做一次标准化，这样就变成了正向指标。

4 实证结果

4.1 基本方程估计

为了检验基本假设，用影响因子做因变量，平均引文数、平均作者数、地区分布数、海外论文比、基金论文比、时效性指数作为自变量，采用普通最小二乘法进行回归，发现海外论文比系数不显著，经检查原始数据发现，我国医学学术期刊的海外论文偏少，537 种期刊中，海外论文比为 0 的期刊有 422 种。原因是多方面的，其中之一就是绝大多数期刊是中文期刊，即使有海外论文，一般也是海外华人撰写的，主要还是语言问题。将海外论文比变量删除再进行回归，结果如表 2 的回归 2 所示。所有变量都通过了统计检验，R^2 值为 0.379，拟合优度较低，说明了平均引文数、平均作者数、地区分布数、基金论文比、期刊时效性 5 个变量解释了影响因子的 37.9%，或者说，这 5 个指标能够提供除了期刊影响因子之外的 62.1%的信息，如果从学术期刊评价的角度而言，为了全面评价学术期刊，这 5 个指标是比较重要的。除了假设 6 没有得到验证外，其他假设都得到了验证。

表 2　普通最小二乘法回归结果

自变量	C	YW	ZZ	DQ	HW	JJ	SX	R^2	n
回归 1	-0.846***	0.006**	0.096***	0.012***	-0.006	0.152**	0.004***	0.387	537
	(-11.649)	(2.330)	(7.618)	(7.573)	(-0.032)	(2.531)	(9.635)		
回归 2	**-0.837*** **	**0.007*** **	**0.095*** **	**0.012*** **	**--**	**0.148** **	**0.004*** **	**0.379**	**537**
	(-11.475)	**(2.686)**	**(7.465)**	**(7.655)**	**--**	**(2.481)**	**(9.435)**		

4.2 分位数回归

为了进一步分析期刊影响因子在不同水平下受其他因素影响的特点，将影响因子分为 10 个分位（τ=0.1～0.9），采用分位数回归进行估计。由于数据量较大，采用平滑算法进行估计，结果如表 3 所示。随着 τ 值变大，拟 R^2 由 0.179 逐渐提高到 0.276，也就是说，5 个自变量对高影响因子期刊的解释程度要好于低影响因子期刊。

表 3 分位数回归结果

自变量	C	YW	ZZ	DQ	JJ	SX	拟 R^2
τ=0.1	-0.420***	-0.005	0.051***	0.006***	0.152***	0.002***	0.179
	（-4.759）	（1.574）	（4.180）	（5.197）	（3.204）	（3.136）	

τ=0.2	−0.458***	0.010***	0.047***	0.007***	0.152***	0.002***	0.204
	(−5.793)	(5.517)	(3.395)	(6.060)	(3.375)	(4.106)	
τ=0.3	−0.539***	0.009***	0.061***	0.009***	0.158***	0.002***	0.215
	(−7.193)	(5.133)	(5.080)	(7.246)	(3.496)	(4.848)	
τ=0.4	−0.531***	0.008***	0.058***	0.009***	0.172***	0.002***	0.219
	(−7.272)	(4.302)	(4.807)	(7.175)	(3.291)	(4.987)	
τ=0.5	−0.619***	0.007***	0.056***	0.009***	0.210***	0.003***	0.220
	(−7.187)	(3.097)	(4.559)	(7.325)	(3.488)	(5.095)	
τ=0.6	−0.742***	0.006**	0.072***	0.011***	0.212***	0.004***	0.230
	(−8.088)	(2.057)	4.939	(8.031)	3.060	(7.321)	
τ=0.7	−0.887***	0.006*	0.096***	0.012***	0.204***	0.004***	0.243
	(−10.532)	(1.651)	(6.640)	(8.179)	(2.914)	(9.525)	
τ=0.8	−0.965***	0.007*	0.140***	0.014***	0.050	0.004***	0.254
	(−10.137)	(1.779)	(5.686)	(6.433)	（0.505）	（6.691）	
τ=0.9	−1.025***	0.009	0.134***	0.016***	0.252	0.005***	0.276
	（−10.468）	（1.492）	（4.457）	（4.846）	（1.183）	（5.442）	

图 1 是平均引文数分位数回归的结果，当 τ=0.1 和 τ=0.9 时，平均引文数对期刊影响因子没有影响，也就是说，当期刊影响因子很低或很高时，期刊论文平均引文数对影响因子没有影响，随着 τ 的取值从小到大，平均引文数对影响因子的影响从大到小，当 τ=0.6、τ=0.7 时达到最低，当 τ=0.8 时又缓慢升高。可能原因是低影响因子期刊编辑部在征稿时往往强调论文的引文数量要求，而高影响因子期刊则没有必要刻意要求这一点。

图 2 是平均作者数分位数回归的结果，当 τ≤0.5 时，曲线基本是一条水平线，平均作者数对影响因子的影响基本稳定不变，而当 τ>0.5 时，曲线直线上升，平均作者数对影响因子的影响越来越大，说明越是高质量的论文，论文作者数所起的作用越大。

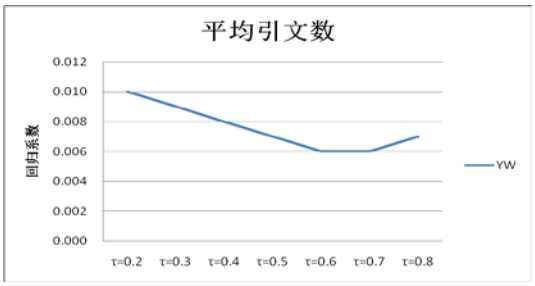

图 1 平均引文数分位数回归

图 2 平均作者数分位数回归

图 3 是地区分布数分位数回归的结果，除了 τ=0.3 和 τ=0.4 时回归系数相等外，曲线基本是一条上升曲线，说明影响因子越高的期刊，其论文作者地区分布数对影响因子的影响越大。

图 4 是期刊时效性指标分位数回归的结果，呈现一种梯形结构，当 τ≤0.4 时，期刊时效性对影响因子的影响不变，或者说，低影响因子期刊的时效性对影响因子的影响稳定。当 0.6≥τ>0.4 时，曲线直线上升，说明对于中等影响因子的期刊，其时效性对影响因子的影响随着影响因子的提高而提高。当 0.8≥τ>0.6 时，期刊时效性对影响因子的影响又保持不变。当 τ>0.8 时，影响因子越高的期刊，其受时效性的影响越大。

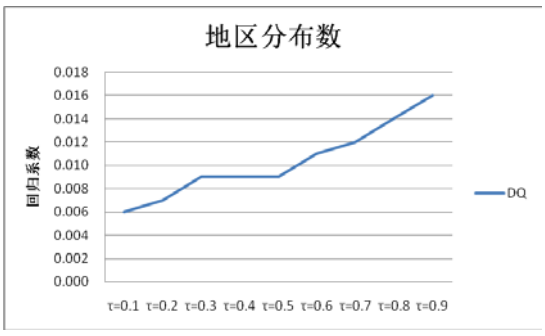

图 3 地区分布数分位数回归

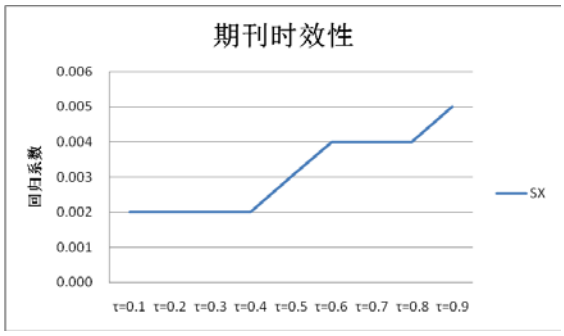

图 4 期刊时效性分位数回归

图 5 是基金论文比分位数回归结果，整体呈现缓慢上升曲线，变化不大，但是当 τ=0.8 和 τ=0.9 时，基金论文比对期刊的影响因子没有影响，或者说，影响因子较高期刊的影响因子不受基金论文比的影响，而影响因子中等偏上及以下期刊，其影响因子受基金论文比的影响比较稳定。

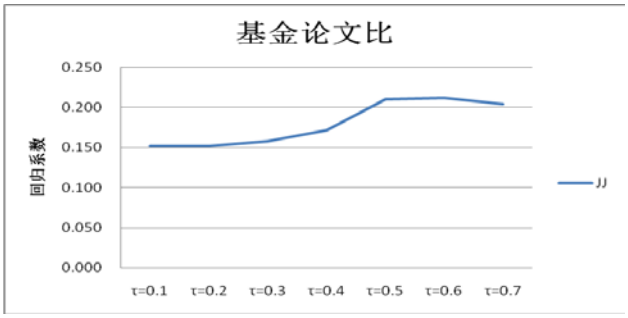

图 5 基金论文比分位数回归

5 结论与讨论

5.1 分位数回归有利于详细分析影响因子处于不同档次的期刊之影响因素的特点

传统最小二乘法回归只能从宏观上分析期刊影响因子的影响因素，本质上是一种均值回归，而且当面临数据较少、数据分布异常等问题时处理比较复杂，而分位数回归可以分析期刊影响因子在不同水平上的影响因素状况，能提供更为详细的信息。将二者相结合，既可以了解期刊影响因子影响因素的宏观图景，也可以了解微观状况。

5.2 影响因子影响因素分析有助于期刊评价指标选取

在采用指标体系评价期刊时，影响因子影响因素分析可以作为选取学术期刊评价指标的一种辅助方法。主要看拟合优度，当 R^2 较低时，说明基金论文比、平均作者数、地区分布数等自变量对影响因子的解释程度低，或者说，这些自变量能够提供除了期刊影响力以外的其他重要信息，而当 R^2 较高时（比如大于 0.8），就要慎重处理，因为指标间重复量？很大，对数据处理和评价方法的要求也随之提高。当然，最重要的还是要从理论上分析该指标是不是适合用来进行期刊评价。

5.3 对于不同水平期刊的评价，评价方法可能不同

这个问题似乎较少有人意识到。传统上采取多属性评价方法对学术期刊进行评价，往往采取一套指标体系，采取同样的方法进行评价。本文的研究发现，高影响因子和低影响因子期刊，平均引文数与期刊影响因子无关；高影响因子期刊，基金论文比与影响因子无关。如果是对高质量期刊进行评价，比如国家期刊评优，是否有必要选取平均引文数和基金论文比指标就值得商榷，虽然对于绝大多数期刊而言，平均引文数和基金论文比与影响因子是相关的，而少部分优秀期刊，这种相关关系并不存在。至少可以说，对不同水平的期刊进行评价，即使指标选取相同，其权重和评价方法也有可能不同。这样的道理对于其他评价对象也是适

用的。比如在大学评价时，若只评价世界一流高校，则诺贝尔奖得主人数也许是个合适的指标；若对所有高校都加以评价，仍采用“诺贝尔奖得主人数”指标，就明显不合适。

参考文献

[1]SCHUBERT A., GLANZEL W., Braun T. Relative citation rate:A new indicator for measuring the impact of publications. In: D.Tomov, L. Dimitrova (eds),Proceeding of the first national conference with international participation on scientometrics and linguistics of scientific text, Varna[M]. 1983,PP80-81

[2]VINKLER P. Evaluation of some methods for the relative assessment of scientific publications[J]. scientometrics,1986(10):157-177

[3]BRAUN T., GLANZEL W. World flash on basic research. A topographical approach to world publication output and performance in science[J].1990,Scientometrics,19:159-165

[4]MOED. H. F.,R.E. DE BRUIN.. New bibliometric tools for the assessment of national research performance[J]. Scientometrics,1995,33:381-422

[5]Hirsch,J.E. An index to qualify an individual’s scientific research output. Proceeding of the national academy of sciences USA[M]. 2005,102:16569-16572

[6]于挨福、马虎兆. 科学、科学研究类期刊影响因子相关因素的实证研究[J]. 科学学研究，2008(8)：767-772

[7]姜春林. 期刊 h 指数与影响因子之间关系的案例研究[J]. 科技进步与对策，2007（9）：78-80

[8]刘雪立、董建军、周志新. 我国医学期刊出版周期与即年指标关系的调查研究[J]. 中国科技期刊研究，2007（4）：597-599

[9]金碧辉、汪寿阳、任胜利等. 论期刊影响因子与论文学术质量的关系[J]. 中国科技期刊研究，2000(11)：202-205

[10] Koenker R, Gilbert B. Regression quantiles [J]. Econometrica, 1978, 46(1):33-50.

[11]Roger Koenker, Kevin F. Hallock. Quantile Regression [J]. Journal of Economic Perspectives, 2001, 15(4):143-156.

[12]Bernd Fitzenberger, Peter Winker. Improving the computation of censored quantile regressions [J]. Computational Statistics & Data Analysis, 2007, 52: 88-108.