doi:10.3772/j.issn.2095-915x.2023.05.005

作者学术经验与被引频次的关系探讨

张丽华[1,2] 姚长青[1]

1. 中国科学技术信息研究所 北京 100038；
2. 山西财经大学 太原 030006

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 开展作者学术经验与论文被引频次的关系研究，有助于挖掘两者之间的影响关系，丰富论文学术影响力理论研究。[方法 / 过程] 以“高被引科学家数据库”中的 194439 名科学家为研究对象，选择 Kruskal-Wallis 秩和检验方法进行单因素方差分析，判断独著作者、第一作者和末位作者的论文被引频次是否存在差异。以 aa、h 指数、h_m 指数为自变量，以总被引频次为因变量，并选择发文量、第一作者发文量、研究多样性等 11 个控制变量，采用负二项回归模型探讨作者学术经验与论文被引频次的关系。[局限] 研究对象较单一，仅选择高被引科学家进行研究。[结果 / 结论] 当作者担任不同角色时，论文被引频次之间存在显著差异。使用负二项回归分析作者学术经验与被引频次关系时发现，无论自变量采用 aa、h 指数还是 h_m 指数，其回归系数均是一个较大的正数，说明作者学术经验确实对论文被引频次产生了积极的影响。

关键词：学术经验；论文被引频次；负二项回归；论文影响力

中图分类号：G35; G316

Research on the Relationship Between Author’s Academic Experience and Citation Counts

ZHANG Lihua[1,2] YAO Changqing[1]

1. Institute of Scientific and Technical Information of China, Beijing 100038;
2. Shanxi University of Finance and Economics, Taiyuan 030006

Abstract: [Purpose/Signficance] The research on the relationship between authors’ academic experience and citation counts of papers is helpful to explore the relationship between them and enrich the theoretical research on academic influence of papers. [Methods/Processes] 194439 scientists in the “Database of highly cited scientists” are selected as research objects. Kruskal-Wallis rank sum test is used for one-way analysis of variance to determine whether there is a difference in citation counts between single author, first author and last author. With aa, h index and h_m index as independent variables and total citation counts as dependent

基金项目 中国博士后科学基金资助项目“研究评估中科学计量学的不确定性现象及应对策略研究”（2018M631551）。

作者简介 张丽华（1986-），博士，副教授，研究方向为科学计量学与科研评价，E-mail：happy2004zlh@163.com；姚长青（1974-），博士，研究员，研究方向为情报理论与方法。

引用格式 张丽华 , 姚长青 . 作者学术经验与被引频次的关系探讨 [J]. 情报工程 , 2023, 9(5): 59-72.

doi:10.3772/j.issn.2095-915x.2023.05.005

variable, 11 control variables, including the number of published papers, the number of published papers authored by first author and the diversity of research are selected, and negative binomial regression model is used to explore the relationship between authors' academic experience and citation counts.[Limitations] The object of the study is relatively simple, and only highly cited scientists are selected. [Results/Conclusions] There are significant differences in citation counts when authors played different roles. When using negative binomial regression to analyze the relationship between the author's academic experience and citation counts, it is found that the regression coefficient is a large positive number no matter the independent variable adopts aa, h index or h_m index, indicating that the author's academic experience does have a positive impact on the citation counts of the paper.
Keywords: Academic Experience; Citation Counts; Negative Binomial Regression Model; Impact of Paper

引言

论文被引频次预测是科学计量学的热门研究话题。面向预测的论文被引频次影响因素研究已积累了丰富的研究成果。这些影响因素大致 4 类：（1）论文相关因素。包括标题长度、论文长度、论文主题、参考文献数量、文献类型、参考文献多样性、是否受基金资助、是否开放获取等。（2）作者相关因素。包括学术经验、性别、年龄、国籍、隶属机构、作者人数、合作、h 指数、累积被引量、作者发文量等。（3）期刊相关因素。包括期刊影响因子、期刊发文量、期刊总被引量、期刊语言等。（4）其它。包括出版时间、论文下载量、社交媒体转发、评论等。

本研究主要关注作者相关因素中的“作者学术经验”。由于研究人员之间的天赋存在较大差异，导致学术经验和论文被引频次之间的关系并不十分密切，但仍然存在这样的共识：作者学术经验很重要 [1]。Sun 等 [2] 发现作者对研究工作的影响比机构更大。作为一个预测论文被引频次的关键因素，作者学术经验测度指标主要包括 4 种：（1）发文量指标。如作者在目标论文之前已发表的论文数量。（2）引用数指标。如作者在目标论文之前已经获得的总被引次数、作者在目标论文之前已发表论文的平均被引频次。（3）学术年龄指标。学术年龄等于学者最新论文的发表年份减去第一篇论文的年份之差加 1。（4）作者状态指标。如作者是否为高被引学者。（5）综合指标。如作者在目标论文之前的 H 指数、第一作者的 H 指数等。

已有部分研究从论文产出角度探索了学术经验与论文被引频次之间的关系。Hanssen 等 [1] 使用作者发文量测度学术经验，发现学术经验对论文被引频次有显著的正向影响，但是这种影响会随着经验水平的提高而迅速减弱。也就是说，年轻的研究人员能够相对较快地学会高水平研究所需的技能与知识。有经验的研究人员将产出更高质量的研究，并最终导致该研究被更频繁地引用。Walters[3] 使用第一作者在 2001—2002 年的发文量、第一作者在 2001—2002 年发文量的被引用次数，以及第一作者 2001—2002 年论文被引用次数除以发文量 3 个指标测度作者的学术经验，并探讨学术经验与论文被引频次的关系，结果表明第 3 个指标对因变量论文被引用次数具有显著影响。Peng 和 Zhu[4] 发现第一作者发文史是论文被引频次的重要预测因素，其中发文史是指论文发表年与数据检索年之间的时间差。相反，Ruan 等 [5] 认为第一作者发文史对单篇论文发表 5 年后的被引频次预测贡献很小。

还有些学者从论文被引角度分析作者经验与论文被引频次的关系。Dalen 等 [6] 使用作者累积被引频次来表示作者学术经验，结果发现作者学术经验是论文被引用次数的最佳预测因素。Bornmann 和 Daniel[7] 研究了作者状态（author’s status，即作者是否为 ISIHighlyCited.com 收录的高被引学者）对论文被引频次的影响，结果发现，作者中高被引学者越多，则论文的被引频次越高。Fu 和 Aliferis[8] 在测度学术经验时使用了第一作者的发文数、第一作者的累积被引用次数、末位作者的发文数和末位作者的累积被引用次数 4 项指标，分析其与论文被引频次的关系，结果表明末位作者的累积被引用次数与第一作者的累积被引用次数指标对论文被引用频次有显著正向影响。程子轩等 [9] 在构建学术论文被引频次预测模型时，选择了作者数量、作者 h 指数、作者发文量、作者论文的被引频次 4 类 10 个作者特征指标，经过相关分析与逐步回归发现，代表作者学术经验的第一作者发表论文的篇均被引量指标能够很好地预测论文被引频次。Wang 等 [10-11] 发现，第一作者在目标论文之前的 h 指数是影响引用的关键因素。Abramo 等 [12] 分析了高产作者与高被引作者之间的关系，发现两者中度相关，高产作者与生产高被引论文的概率存在一定相关性。同时，大约一半高被引论文的作者是发文量排名前 10% 的作者。Danell[13] 使用作者已发表论文数和已发表论文的引用率定量测度作者过往记录，分析这两个指标能否预测论文的影响力。结果表明，已发表论文的引用率是论文影响力的重要预测因素。相反，已发表论文数反而不重要。Ruan 等 [5] 发现第一作者的总被引频次、篇均被引频次、H 指数均不是预测论文被引频次的重要指标。

实际上，在分析作者学术经验与论文被引频次关系时，不同学者得出了不同的结论。Onodera 和 Yoshikane[14] 总结了影响论文被引频次的不同因素，本文截取了与作者学术经验相关的因素（表 1），从表 1 中可以看到，作者发文量、作者被引量、作者状态这 3 个表示作者经验的指标与论文被引频次的关系在不同的研究中得出不同的结论。这一方面表明当同一主题的不同研究结论存在差异时，应关注不同研究对象、视角与方法的差异，另一方面也表明本文

表 1　影响论文被引频次的作者相关因素

研究	作者发文量 Author’s productivity	作者被引量 Author’s citedness	作者状态 Other status of author
Bornmann 和 Daniel, 2008[7]			A
Fu 和 Aliferis, 2001[8]	C	A	
Haslam 等 2008[15]	A		
He, 2009[16]		B	C
Peng 和 Zhu, 2012 [4]			A
Peters 和 van Raan, 1994[17]	A		
Stewart, 1983[18]		A	C
Van Dalen 和 Henkens, 2001[6]		A	
Walters, 2006[3]		A	

注：A- 强预测因素或明确预测因素；B- 弱预测因素或预测能力取决于模型；C- 不显著或负向预测因素

doi:10.3772/j.issn.2095-915x.2023.05.005

能够在现有研究基础上丰富作者经验与论文被引频次之间的关系研究。

同时，在现有的作者学术经验指标中，除了第一作者，较少考虑其他作者角色，如末位作者、单一论文作者以及第一作者与单一作者论文分别对应的论文数与被引用次数。此外，已有研究多是从单篇论文层面关注作者学术经验与论文被引频次之间的关系，这表现为被解释变量通常是单篇论文的被引频次，解释变量则为该篇论文所对应的特征，如作者人数、期刊影响因子、作者在这篇论文之前的发文量等。较少从作者层面，尤其是作者整个职业生涯所有学术论文的汇聚层面探讨学术经验与论文被引频次之间的关系。

因此本研究旨在从作者层面探讨作者学术经验与论文被引频次的关系。我们选择斯坦福大学 John P. A. Ioannidis 教授团队发布的“高被引科学家数据库”中的 194439 名科学家为研究对象，主要解决以下 2 个问题：（1）当作者担任不同角色时，学术经验与论文被引频次之间的关系是否存在差异？（2）作者层面作者经验与论文被引频次呈现出怎样的关系？通过对这 2 个问题的回答，有助于更深入地探讨学术经验与论文被引频次的关系，使研究机构在聘用、晋升科研人员时合理考虑科研人员的学术经验。

1 数据来源与方法

2022 年 11 月 3 日，斯坦福大学 John P. A. Ioannidis 教授团队发布开放获取的 2022 年“高被引科学家数据库”（第五版，https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/5）。该数据库 2019 年 7 月 6 日发布了第 1 版，每年更新 1 版，第 2 版和第 3 版的发布时间分别是 2020 年 10 月 8 日和 2021 年 10 月 19 日。2022 年的情况较特殊，分别于 10 月 10 日和 11 月 3 日更新了第 4 版和第 5 版。这两版的主要区别在于研究子领域的划分方法与数量，其余计量指标没有发生变化。该团队指出，第 5 版更合适，应该取代第 4 版。

“高被引科学家数据库”包括“年度影响力数据集”（single recent year dataset）和“职业生涯影响力数据集（1960—2022）”（career-long database）两个排名。通过遴选出基于 c 值（c-score）或子领域排名前 2% 的前 10 万名科学家，来自 22 个学科领域及 174 个子领域的 200196 名科学家入选“2022 年度影响力数据集”，194983 名科学家入选“职业生涯数据集”。本研究通过对数据集的清洗，最终选择第 5 版“职业生涯数据集”中的 194439 名科学家为研究对象。数据集中包括丰富的计量指标信息，包括科学家姓名、机构、国家、发表第一篇论文年份、最近一篇论文年份、总被引频次、h 指数、h_m 指数、作者排名最高的领域、子领域等。

本研究拟解决的第 1 个问题是当作者担任不同角色时，论文被引频次是否存在差异。根据署名位置，作者角色可以划分为独著、第一作者、末位作者以及除此之外的其他作者。其中，独著作者说明研究工作的构思、设计、分析与论文撰写全部由作者一人完成，作者是论文的全部贡献者。第一作者是合著论文的主要贡献者，他 / 她不仅应该是课题主要观点的拥有者，而且除特殊情况外还必须是科研课题的具体操作者和文章的主要执笔者 [19]。末位作者一般是高级作者 [20]，为“指导、监督和保证所报道作

品的真实性”以及“对作品的科学准确性、有效方法、分析和结论承担责任”的个人[21]。当然，在论文作者署名完全按照贡献度大小排序时，末位作者也可能对论文的贡献度最小。其他作者是指非独著、非第一、非末位作者的其他作者。其他作者的贡献度一般按照署名位次依次减小。本研究选择对论文做出重要贡献的独著作者、第一作者和末位作者进行研究。

我们采用方差分析方法观察同一作者扮演不同角色时被引频次是否存在差异。首先为194439名科研人员根据角色不同建立3组数据，分别是独著作者组、第一作者组和末位作者组。接下来依次进行不同组及组与组之间的正态性、方差齐性检验，根据检验结果选择合适的假设检验方法，本文中选择Kruskal-Wallis秩和检验方法进行单因素方差分析，最后根据分析结果得出不同组被引频次是否存在差异的结论。

本研究拟解决的第2个问题是探讨作者层面学术经验与被引频次的关系。论文被引频次服从偏态分布，泊松回归和负二项回归是针对偏态分布的常用模型。但泊松分布要求数据的总体方差等于均值，一般现实世界的数据较难满足这一要求。因此，本研究采用负二项回归分析作者经验与论文被引频次的关系。同时，负二项回归也是其他学者研究这两个变量关系的常用方法[6-7,14,22]。本研究在使用负二项回归模型时采用了马萨诸塞大学阿默斯特分校（University of Massachusetts, Amherst）Sachin Date的研究：https://timeseriesreasoning.com/contents/negative-binomial-regression-model/。负二项回归的基本操作步骤为：（1）对数据集进行泊松回归拟合，获得拟合率向量λ；（2）对数据集进行aus OLS回归拟合，获得α的值；（3）使用第2步中获得的α对数据集进行负二项回归拟合；（4）使用拟合的负二项回归模型进行预测；（5）检验负二项回归模型的拟合优度。

进行负二项回归分析采用的指标见表2。学术经验我们选择了3个指标，学术年龄、h指数和h_m指数。学术年龄和h指数是经典的用于衡量作者经验的指标。h_m指数[23]主要用于h指数在多作者论文中存在不公平的问题，是利用论文平均作者数量标准化之后的指标。其计算方式类似于h指数，科学家的h_m篇论文至少被引用了h_m次，其他论文的被引次数都少于h_m。只不过在计算论文数量时，将论文数量除以该篇论文的作者人数。例如一篇论文有3名作者，则对每名作者来说，其发文量等于1/3。此外，在作者层面，计算引用影响指标时排除自引更合理[24]。因此本研究中涉及的所有引用指标均排除了作者自引。

选择控制变量时主要考虑了以下方面：（1）作者发文量。一般来说，作者发表的论文数量越多，则其总被引频次可能越高。发文量是影响作者经验与被引频次关系的重要因素。（2）作者角色的差异。当作者处于不同角色时，其对研究的贡献也存在很大的差异。如第一作者是研究的最大贡献者，而末位作者可能是论文的通讯作者，也可能对研究的贡献最小。根据署名位置，作者角色可以划分为独著、第一作者、末位作者、其他作者（指除独著、第一作者、末位作者之外处于其他署名位置的作者）。本研究考察了作者在其职业生涯中，处于第一作者、末位作者等不同角色时的发文量与被引频次的关系。（3）研究的多样性。本研究主要

doi:10.3772/j.issn.2095-915x.2023.05.005

从施引文献的角度考虑研究的多样性，不同的施引文献数量越多，可以认为该研究涉及的主题越多样。多样的研究主题可能为研究带来更多的引文。（4）Scopus 停止收录的论文数量与引用频次。期刊会因不当出版行为被 Scopus 停止收录。停止收录可能说明期刊中的论文存在质量问题。通过停止收录论文与引文分析，可以清楚地了解作者被引用频次的来源。

表 2　学术经验与论文被引频次指标

类型	指标	缩写	描述
解释变量	学术年龄	aa	作者最新论文的发表年份与第一篇论文的年份之差加 1
	h 指数	h	科学家的 h 篇论文至少被引用 h 次，其他论文的被引次数都少于 h，截止到 2021 年底作者的 h 指数
	h_m 指数	h_m	利用论文作者人数对 h 指数进行标准化，截止到 2021 年底作者的 h_m 指数
被解释变量	总被引频次	tct	作者在 1960—2021 年发表所有论文的总被引频次
控制变量	发文量	np	作者在 1960—2021 年的总发文量
	独著论文数量	sp	作者在 1960—2021 年发表的独著论文数量
	独著论文被引频次	spc	作者在 1960—2021 年发表的独著论文的被引频次
	第一作者论文数量	fp	作者在 1960—2021 年以第一作者身份发表的论文数量
	第一作者论文被引频次	fpc	作者在 1960—2021 年以第一作者身份发表的论文的被引频次
	末位作者论文数量	lp	作者在 1960—2021 年以末位作者身份发表的论文数量
	末位作者论文被引频次	lpc	作者在 1960—2021 年以末位作者身份发表的论文的被引频次
	未被引用的论文数量	ncp	作者在 1960—2021 年未被引用的论文数量
	研究的多样性	dr	作者在 1960—2021 年论文的不同施引论文的数量
	Scopus 停止收录的论文数量	stp	作者在 1960—2021 年被 Scopus 停止收录的论文数量
	Scopus 停止收录论文的被引用频次	stpc	作者在 1960—2021 年被 Scopus 停止收录的论文的被引用频次

由于自变量和控制变量之间的原始数据差别较大，因此对其进行标准化处理。标准化方法采用最大最小标准化方法（min-max normalization）。这种方法简单易理解，不改变数据分布，采用的公式为

$$Y=(X-X\_\min)/(X\_\max-X\_\min)$$

其中，Y 是指标的标准化值；X 为指标的原始值；$X\_\max$ 与 $X\_\min$ 分别对应指标的最大值和最小值[25]。

本研究中所使用的数据分析工具为 EXCEL，SPSS 和 Python。

2　研究结果

2.1　变量描述性统计

对本研究中涉及的 15 个变量进行描述性统计分析，见表 3。从表 3 中可以看出，tct 的离散程度较高，说明不同作者在 1960-2021 年发表论文的总被引频次差别较大。3 个自变量中，aa 主要考查科研人员的职业生涯长度，均值为 35，最长的职业生涯与最短的职业生涯相差 68 年。h_m 指数与 h 指数相比，其均值与方差都更小，这与 h_m 指数的计算方法有关，其是在 h 指数的

基础上对论文平均作者人数的标准化。

在控制变量中，第一作者（fp,fpc）、末位作者（lp,lpc）的发文量和被引量均值高于独著作者（sp,spc）。这与研究中合作现象越来越普遍的趋势一致。Dr 是研究的多样性指标，也是一个均值和方差都很大的指标。ncp 是未被引用的论文数量，其均值是总发文量均值的15.9%。这说明从平均水平来看，相比于至少被引用 1 次的论文而言，高被引科学家群体未被引用的论文数量在职业生涯的总发文量中占少数。Stp,stpc 是从未被 Scopus 数据库停止收录的角度考察其对被引频次的影响，总体来看，论文被 Scopus 停止收录现象较不常见，但可以从另一个独特视角考察科学家的影响力情况。

表 3　变量的描述性统计

指标	样本量	最小值	最大值	均值	方差
tct	194439	51	428620	8114.58	116964535.827
aa	194439	2	70	35.01	122.342
h	194439	3	284	38.16	348.967
h_m	194439	0.836	114.999	18.045	59.931
np	194439	2	3791	199.01	26254.871
sp	194439	0	2234	14.47	670.873
spc	194439	0	164371	401.07	1581574.722
fp	194439	0	1368	31.72	892.569
fpc	194439	0	206139	1384.95	4977706.631
lp	194439	0	2360	70.51	6155.000
lpc	194439	0	194533	2376.42	17711277.148
dr	194439	49	315584	6108.15	61496498.744
ncp	194439	0	2196	31.71	1599.545
stp	194439	0	300	.69	16.514
stpc	194439	0	8867	71.19	15649.557

2.2　方差分析结果

方差分析用于解决不同角色的作者发文量和被引频次是否存在差异的问题。应用方差分析需要满足 3 个条件：（1）各样本相互独立；（2）各对比组资料服从正态分布;（3）各总体方差相等，即方差齐 [26]。高被引科学家数据集中各个科学家个体相互独立，满足条件 1。我们使用 Q-Q 图来检验 sp,fp,lp,spc,fpc 和 lpc 指标是否服从正态分布（图 1、图 2）。Q-Q 图是根据样本数据的分位数与理论分布（如正态分布）的分位数的符合程度绘制的。如果实际数据服从正态分布，则所有分位数应该落在截距为样本均值，斜率为样本标准差的直线上。从图 1、图 2 可以看出，不同作者角色的发文量与被引频次指标并不服从正态分布。

doi:10.3772/j.issn.2095-915x.2023.05.005

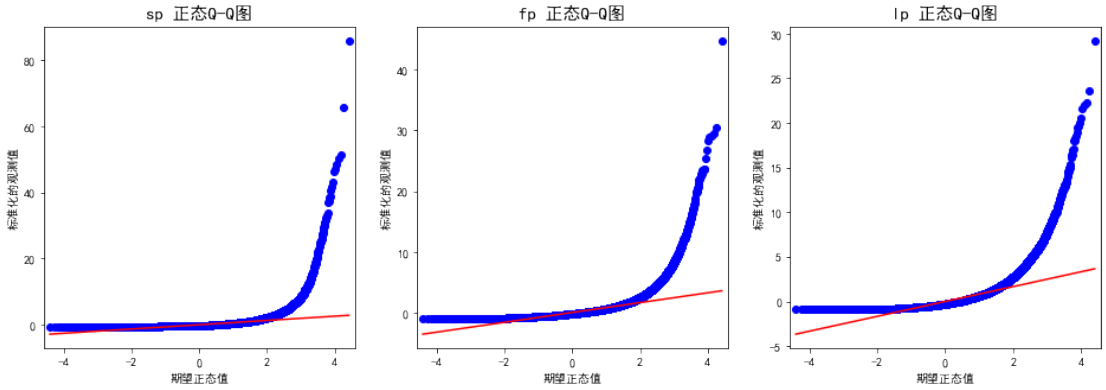

图 1　作者角色发文量指标 Q-Q 图

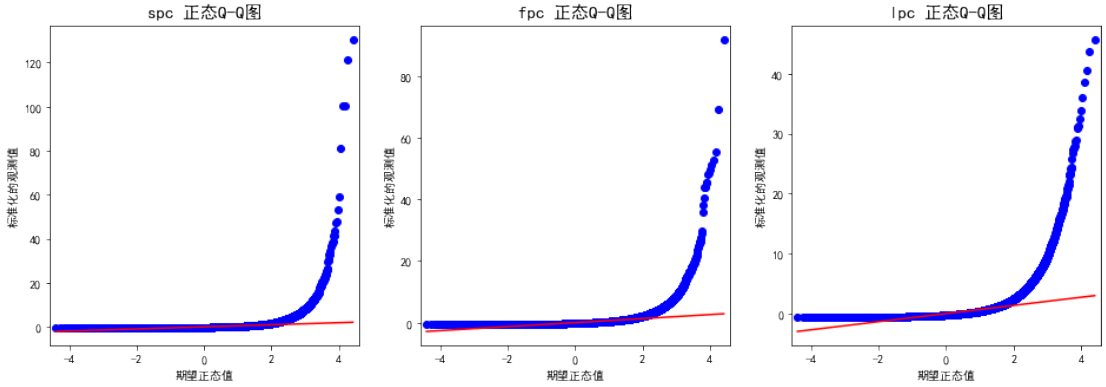

图 2　作者角色被引量指标 Q-Q 图

本研究通过绘制残差图来检验方差齐性（图 3、图 4）。如果拟合值和残差的散点随机分布在一个水平带之内，没有离群点，而且其离散程度基本上一样，表示满足方差齐性的要求。从图 3、图 4 中可以看出，作者角色的发文量和被引频次指标不满足方差齐性的前提要求。

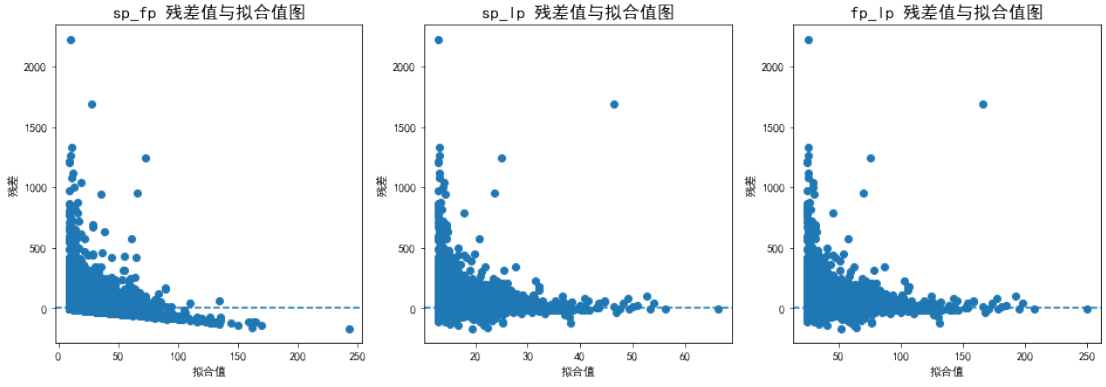

图 3　作者角色发文量指标的残差图

因作者角色相关的 6 个指标不满足正态性和方差齐性的前提条件，本研究选择 Kruskal-Wallis 秩和检验进行单因素方差分析。经过计算，sp,lp 和 fp 的统计量 H 等于 196993.5，p 值为 0.0，说明不同作者角色发文量不全相等。为了找出 sp 与 lp、sp 与 fp、lp 与 fp 之间究竟是哪两个均值不相等，我们采用 Tukey HSD 方法进行多重比较。结果表明，sp,lp 与 fp 的均值

两两之间均存在显著差异。由此我们可以得出结论：同一作者在扮演不同角色时其发文量存在显著差异。采用相同的方法，我们对 spc,lpc 和 fpc 进行 Kruskal-Wallis 检验，统计量 H 等于 179298.9，p 值为 0.0。多重比较也显示其均值两两之间存在显著差异。因此，同一作者在扮演不同角色时其发表论文的被引量也存在显著差异。

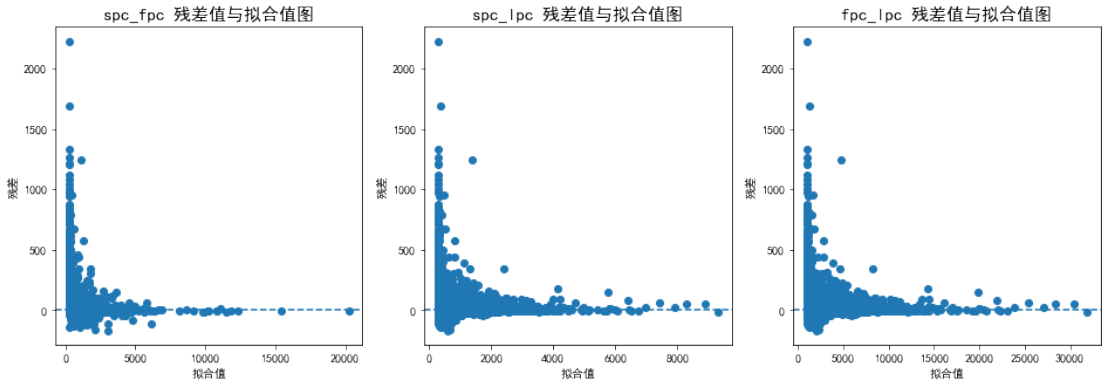

图 4　作者角色被引频次指标的残差图

2.3　负二项回归结果

建立负二项回归模型之前需要判断自变量、控制变量之间是否存在共线性问题。因多数变量不满足正态性要求，此处选择 spearman 相关系数。11 个控制变量、3 个自变量两两之间的相关系数见表 4—表 7。相关系数大于 0.6 以蓝色底纹表示。从表 4 中可以看出，除 fp,fpc,stp 指标之外，其余控制变量均存在与其他变量相关系数较高的情况。自变量 aa 与所有控制变量均不相关。H 指数与 np,lpc,dr 存在相关关系。而 h_m 指数与 np,lp,lpc,dr 存在相关关系。多个变量之间存在较高相关系数，提示由这些变量建立的回归模型可能存在共线性问题。

表 4　控制变量之间的 spearman 相关系数

	np	sp	spc	fp	fpc	lp	lpc	dr	ncp	stp	stpc
np	1	.075**	-.188**	.504**	.231**	.809**	.588**	.602**	.761**	.246**	.518**
sp	.075**	1	.674**	.242**	-.044**	.032**	-.022**	-.076**	.247**	-.069**	-.086**
spc	-.188**	.674**	1	-.010**	.036**	-.137**	-.024**	-.065**	-.151**	-.164**	-.156**
fp	.504**	.242**	-.010**	1	.461**	.279**	.074**	.152**	.498**	.150**	.216**
fpc	.231**	-.044**	.036**	.461**	1	.096**	.265**	.530**	-.006*	-.019**	.329**
lp	.809**	.032**	-.137**	.279**	.096**	1	.770**	.481**	.609**	.191**	.410**
lpc	.588**	-.022**	-.024**	.074**	.265**	.770**	1	.726**	.243**	.050**	.482**
dr	.602**	-.076**	-.065**	.152**	.530**	.481**	.726**	1	.207**	.048**	.657**
ncp	.761**	.247**	-.151**	.498**	-.006*	.609**	.243**	.207**	1	.268**	.322**
stp	.246**	-.069**	-.164**	.150**	-.019**	.191**	.050**	.048**	.268**	1	.313**
stpc	.518**	-.086**	-.156**	.216**	.329**	.410**	.482**	.657**	.322**	.313**	1

**. 在 0.01 级别（双尾），相关性显著。*. 在 0.05 级别（双尾），相关性显著。

表 5　自变量 aa 与控制变量之间的 spearman 相关系数

	np	sp	spc	fp	fpc	lp	lpc	dr	ncp	stp	stpc
aa	.222**	.423**	.278**	.186**	-.038**	.249**	.266**	.147**	.197**	-.078**	.036**

**. 在 0.01 级别（双尾），相关性显著。*. 在 0.05 级别（双尾），相关性显著。

doi:10.3772/j.issn.2095-915x.2023.05.005

表 6 自变量 h 与控制变量之间的 spearman 相关系数

	np	sp	spc	fp	fpc	lp	lpc	dr	ncp	stp	stpc
h	.685**	-.079**	-.085**	.223**	.505**	.562**	.758**	.903**	.228**	.062**	.575**

**. 在 0.01 级别（双尾），相关性显著。*. 在 0.05 级别（双尾），相关性显著。

表 7 自变量 h_m 与控制变量之间的 spearman 相关系数

	np	sp	spc	fp	fpc	lp	lpc	dr	ncp	stp	stpc
h_m	.622**	.236**	.243**	.311**	.400**	.612**	.718**	.647**	.243**	.045**	.403**

**. 在 0.01 级别（双尾），相关性显著。*. 在 0.05 级别（双尾），相关性显著。

为了解决共线性问题，采用逐步回归进行变量筛选。逐步回归的基本思想将全部自变量按其对因变量的影响程度大小，从大到小地依次把自变量引入方程。每引入一个自变量，就要对它作检验，有统计学意义才引入。当新的自变量进入方程后，就对方程中当时所含有的全部自变量进行检验，一旦发现不具有统计学意义的自变量就立即剔除 [26]。经过回归分析并结合变量的实际意义，本文最后选择如下变量加入负二项回归模型：（1）在以 aa 为自变量的模型中，包括所有控制变量；（2）在以 h 指数为自变量的模型中，控制变量包括 dr,lpc,fpc,lp,np,fp,spc,ncp,stpc；（3）在以 h_m 指数为自变量的模型中，控制变量包括 dr,lpc,fpc,np,lp,ncp,fp,spc,sp。

各模型的回归结果见表 8—表 10。我们为每个自变量构建了两个模型，模型 1 中只包括控制变量，模型 2 中包括控制变量与自变量。分别对两个模型进行检验，并计算其对数似然值。接下来检验 2 个模型对数似然比，χ2 统计量等于模型 1 和模型 2 对数似然值差值的 2 倍。若 $\chi^2 \geq \chi_{0.05}^2$，则拒绝原假设，说明作者经验显著影响被引频次，反之，则说明不存在显著影响。

表 8 作者学术经验 aa 与被引频次负二项回归结果

		tct
模型 1	np	9.1906*
	sp	-1.1809*
	spc	3.3838*
	fp	-4.1792*
	fpc	17.985*
	lp	-1.5156*
	lpc	3.8047*
	ncp	-6.5362*
	dr	25.4707*
	stp	1.2931*
	stpc	-2.4819*
	χ2	32100
	$\chi_{0.05}^2$	19.675
模型 2	**np**	9.4142*
	sp	-3.6948*
	spc	2.6603*
	fp	-4.2744*
	fpc	18.9153*
	lp	-1.8482*
	lpc	3.4966*
	ncp	-6.8418*
	dr	25.2612*
	stp	0.392*
	stpc	-2.3964*
	aa	0.7839*
	χ2	32500
	$\chi_{0.05}^2$	21.026
模型 1.vs. 模型 2	χ^2	1800
	$\chi_{0.05}^2$	3.841

注：“*”代表在 0.05 水平上显著。灰色区域单元格内的数字代表卡方值（χ^2）和卡方临界值（$\chi_{0.05}^2$），非灰色区域单元格内的数字代表回归系数（下同）。

表 9　作者学术经验 h 指数与被引频次负二项回归结果

		tct
模型 1	**np**	9.0692*
	spc	1.6272*
	fp	-3.5183*
	fpc	17.4136*
	lp	-0.6684*
	lpc	3.1588*
	ncp	-8.8408*
	dr	26.0125*
	stpc	**-2.6664***
	$\chi 2$	32600
	$\chi_{0.05}^2$	16.919
模型 2	**np**	-2.3027*
	spc	5.2618*
	fp	-1.1551*
	fpc	8.9309*
	lp	2.8019*
	lpc	-5.5775*
	ncp	1.1891*
	dr	14.0336*
	stpc	0.858*
	h	9.9299*
	χ^2	39300
	$\chi_{0.05}^2$	18.307
模型 1.vs. 模型 2	χ^2	74400
	$\chi_{0.05}^2$	3.841

表 10　作者学术经验 hm 与被引频次负二项回归结果

		tct
模型 1	**np**	9.0756*
	sp	-1.3453*
	spc	3.2292*
	fp	-4.1447*
	fpc	17.8083*
	lp	-1.5168*
	lpc	3.8299*
	ncp	-6.3947*
	dr	24.7532*
	$\chi 2$	32000
	$\chi_{0.05}^2$	16.919
模型 2	**np**	5.4935*
	sp	-7.9665*
	spc	-4.9825*
	fp	-4.7851*
	fpc	12.2746*
	lp	-3.1359*
	lpc	-2.6574*
	ncp	-0.3812*
	dr	26.0647*
	h_m	4.2814*
	χ^2	51700
	$\chi_{0.05}^2$	18.307
模型 1.vs. 模型 2	χ^2	21200
	$\chi_{0.05}^2$	3.841

从表 8—表 10 中我们可以看出：

（1）采用 3 个自变量时，模型 2 的拟合度均优于模型 1。这表现为当自变量为 aa、h 指数和 h_m 指数时，模型 1 与模型 2 对数似然比 $\chi^2 \geq \chi_{0.05}^2$，说明作者学术经验对论文被引频次存在影响。

（2）当自变量为 aa 时，其回归系数为 0.7839，自变量为 h 指数时，其回归系数为 9.9299，而自变量为 h_m 指数时，回归系数为 4.2814。说明作者学术经验与论文被引频次均为正向关系。即作者学术经验越多，则其论文被引频次越高。其次，采用不同指标衡量作者学术经验时，学术经验与论文被引频次呈现的量化关系密切程度有所不同。

（3）在控制变量中，dr 的回归系数在模型中均大于 14，是回归系数的最大值，说明研究的多样性是学术经验与论文被引频次之间关系的最大影响因素。回归系数第二大的指标为 fpc，其值均大于 8，说明其对论文被引频次为正向影响，即第一作者发文量的被引频次指标越大，论文被引频次越高。当自变量使用不同指标时，np 对论文被引频次的影响情况有所不

doi:10.3772/j.issn.2095-915x.2023.05.005

同，当使用 aa 指标与 h_m 指标时，作者总发文量对论文被引频次有正向影响，而使用 h 指数时，则对论文被引频次有负向影响。除此之外，当自变量为 aa 时，第一作者论文数量（fp）和未被引用的论文数量（ncp）对论文被引频次有显著负向影响。当自变量为 h 指数时，独著论文被引频次（spc）对论文被引频次有显著正向影响，而末位论文作者被引频次（lpc）则有显著负向影响。当自变量为 h_m 指数时，独著论文数量（sp），独著论文被引频次（spc）以及第一作者论文数量（fp）越大，则论文被引频次越低。

3 总结与讨论

本文以高影响力科学家为研究对象，从科研人员整体职业生涯的视角探讨作者学术经验与论文被引频次的关系。主要结论如下：

（1）当作者担任独著作者、第一作者和末位作者等不同角色时，论文被引频次之间存在显著差异。从方差分析结果来看，sp,lp 与 fp 的均值两两之间、spc,lpc 和 fpc 均值两两之间均通过了显著性检验。从表 3 中也可以看出，spc、fpc 和 lpc 的均值分别为 401.07、1384.95 和 2376.42。可见末位作者的论文被引用频次更高，其次为第一作者，最后为独著作者。末位作者一般为资历更高的学者。这也验证了“资历更高的学者其论文被引频次更高”的结论[4]。

（2）使用负二项回归分析作者学术经验与被引频次关系时发现，无论自变量采用学术年龄、h 指数还是 h_m 指数，其回归系数均是一个较大的正数，说明作者学术经验确实对论文被引频次产生了积极的影响。作者学术经验积累得越多，其论文被引频次值越大。也就是说，我们从高影响力科学家数据集的角度再次验证了两者之间的正向影响关系。这与 Hanssen 等[27]的研究结论一致。

（3）研究多样性是影响论文被引用频次的最主要因素。本文中研究多样性是指作者在 1960—2021 年论文的不同施引论文的数量。研究多样性对应着科学计量学中更泛化的概念“学科多样性”。最常用的学科多样性指标包括跨领域引用指数、信息熵、布里渊指数和 Rao-Stirling 等[28]。目前学科多样性与论文被引频次的关系尚未有明确结论，仍需进一步深入研究[29]。本文经过分析发现，不论使用哪个自变量，研究多样性 dr 的回归系数在所有自变量和控制变量中均为最大值，说明研究多样性是影响论文被引用频次的最主要因素，该结论进一步丰富了学科多样性与论文被引频次关系理论。

（4）作者学术经验与论文被引频次的量化关系受数据集选择、学术经验计算方法的影响。我们通过负二项回归分析发现，作者学术经验与论文被引频次之间确实存在影响关系。但这种影响关系到底有多大，在使用不同的学术经验指标时呈现出了不同的结果。同时，本研究选择的高影响力科学家数据集，与使用其他数据集分析二者关系时呈现出的结果也不完全相同。因此，我们在分析此类研究主题时，务必仔细选择数据、指标、方法，若要进行对比分析，则应注意数据集、指标、方法的可比性。

未来我们将从以下方面继续开展学术经验与论文被引频次的关系研究：第一，本文研究对象仅涉及高影响力科学家，未来我们将选择

其他科研人员群体与高影响力科学家进行对比分析；第二，本文中作者角色涉及独著作者、第一作者和末位作者，并未涉及除这三种角色之外的其他作者，如在研究中起领导与协调作用的通讯作者，以及人数众多的中间作者，未来将开展作者角色的更细致分析；第三，本文只是从定量的角度分析二者关系，对于二者背后的相互影响机制及深层次原因还需要通过个案分析、定性分析来实现，未来可将定性与定量方法结合以获得更客观的分析结论。

参 考 文 献

[1] HANSSEN T-E S, JØRGENSEN F, LARSEN B. The relation between the quality of research, researchers' experience, and their academic environment[J]. Scientometrics, 2017, 114(3): 933-950.

[2] SUN M, MA T, ZHOU L, et al. Analysis of the relationships among paper citation and its influencing factors: a Bayesian network-based approach[J]. Scientometrics, 2023, 128(5): 3017-3033.

[3] WALTERS G D. Predicting subsequentcitations to articles published in twelve crime-psychology journals: Authorimpact versus journal impact[J]. Scientometrics, 2006, 69(3): 499-510.

[4] PENG T Q, ZHU J J. Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies[J]. Journal of the American Society for Information Science and Technology, 2012, 63(9): 1789-1803.

[5] RUAN X, ZHU Y, LI J, et al. Predicting the citation counts of individual papers via a BP neural network[J]. Journal of Informetrics, 2020, 14(3): 101039.

[6] VAN DALEN H, HENKENS K. What makes a scientific article influential? The case of demographers[J]. Scientometrics, 2001, 50(3): 455-482.

[7] BORNMANN L, DANIEL H D. Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere[J]. Journal of the American Society for Information Science and Technology, 2008, 59(11): 1841-1852.

[8] FU L, ALIFERIS C. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature[J]. Scientometrics, 2010, 85(1): 257-270.

[9] 程子轩 , 张向先 , 郭顺利 . 基于作者特征和期刊特征的学术论文被引频次预测模型构建与分析 [J]. 情报科学 , 2021, 39(3): 179-184, 192.

[10] WANG M, YU G, AN S, et al. Discovery of factors influencing citation impact based on a soft fuzzy rough set model[J]. Scientometrics, 2012, 93(3): 635-644.

[11] WANG M, WANG Z, CHEN G. Which can better predict the future success of articles? Bibliometric indices or alternative metrics[J]. Scientometrics, 2019(119): 1575-1595.

[12] ABRAMO G, CICERO T, D'ANGELO C A. Are the authors of highly cited articles also the most productive ones?[J]. Journal of Informetrics, 2014, 8(1): 89-97.

[13] DANELL R. Can the quality of scientific work be predicted using information on the author's track record?[J]. Journal of the American Society for Information Science and Technology, 2011, 62(1): 50-60.

[14] ONODERA N, YOSHIKANE F. Factors affecting citation rates of research articles[J]. Journal of the Association for Information Science and Technology, 2015, 66(4): 739-764.

[15] HASLAM N, BAN L, KAUFMANN L, et al. What makes an article influential? Predicting impact in social and personality psychology[J]. Scientometrics, 2008, 76(1): 169-185.

[16] HE Z L. International collaboration does not have greater epistemic authority[J]. Journal of the American Society for Information Science and Technology, 2009, 60(10): 2151-2164.

[17] PETERS H P, VAN RAAN A F. On determinants of citation scores: A case study in chemical engineering[J]. Journal of the American Society for

doi:10.3772/j.issn.2095-915x.2023.05.005

Information Science, 1994, 45(1): 39-49.

[18] STEWART J A. Achievement and ascriptive processes in the recognition of scientific articles[J]. Social Forces, 1983, 62(1): 166-189.

[19] 薛镭 . 关于学术论文第一作者的署名问题 [J]. 编辑学报 , 2003, 15(1): 33-34.

[20] 马英敏 , 田文灿 , 曹仁猛 , 等 . 科技论文中通信作者和末位作者重要性的比较 [J]. 中国科技期刊研究 , 2021, 32(7): 910-916.

[21] MCKNEALLY M. Put my name on that paper: reflections on the ethics of authorship[J]. The Journal of thoracic and cardiovascular surgery, 2006, 131(3): 517-519.

[22] WALTERS G D. Predicting subsequent citations to articles published in twelve crime-psychology journals: Authorimpact versus journal impact[J]. Scientometrics, 2006, 69(3): 499-510.

[23] SCHREIBER M. To share the fame in a fair way, hm modifies h for multi-authored manuscripts[J]. New Journal of Physics, 2008, 10(4): 040201.

[24] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(46): 16569-16572.

[25] 张丽华 , 张康宁 , 赵迎光 , 等 . 科研人员职业生涯学术论文相似度及其对被引频次的影响分析 [J]. 情报学报 , 2022, 41(8): 822-831.

[26] 刘桂芬 . 卫生统计学 [M]. 北京 : 中国协和医科大学出版社 , 2003: 56.

[27] HANSSEN T-E S, JØRGENSEN F. The value of experience in research[J]. Journal of Informetrics, 2015, 9(1): 16-24.

[28] WANG J, THIJS B, GLÄNZEL W. Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity[J]. PLoS One, 2015, 10(5): e0127298.

[29] 刘智锋 , 马永强 , 杨金庆 . 引文学科多样性与论文影响力的关系研究 [J]. 情报杂志 , 2020, 39(7): 190-195, 207.