摘 要

关键词：引用分析

Abstract

**Keywords:** Citation

目录

键入章标题(第 1 级)1

键入章标题(第 2 级)2

键入章标题(第 3 级)3

键入章标题(第 1 级)4

键入章标题(第 2 级)5

键入章标题(第 3 级)6

**未找到图形项目表。**

**未找到图形项目表。**

第一章 绪 论

# 1.1引言

从历史的发展规律中可以看到，世界经济的发展，特别是发达国家，经济发展的主要动力来自于科技创新和技术进步。进入新世纪以来，科技发展速度越来越快，最直接的体现就是科研投入和产出的增加。根据国家统计局的数据（图1.），2012年，我国共投入研究与试验发展经费(Research & Development)经费10298.4亿元，比上年增加了1611.4亿元，增长率为18.5%，其投入强度（与国内生产总值之比）为1.98%，较上年的1.84%有所提高。而R&D经费和2003年的1539.6亿元相比，翻了6倍。而且，以后的投入也会继续增加。同时，科研产出量增加也十分显著，仅以微生物领域来看，进入新世纪以来，我国被SCI收录的论文增长了将近20倍。然而，在看似如此迅速的增长中，如此庞大的科研投入，其中也有一些投入的不合理之处，我国的科研创新效果并不明显。在这样的背景下，如何合理地进行科研投入，如何提高科研投入效率，成了一个亟待解决的问题。若想要解决这个问题，则首先要找到合适的投入方向，如何通过科学评价方法发掘适于科研投入的领域和人才，成为了最直接的问题。

图1.我国的R&D投入及科研产出

在目前的研究中，多是用单一的评价指标，例如*h*指数、影响因子等进行评价，但是仅仅利用单一的评价指标无法得到全面完善的结果，而且这些结果也都是在对已完成的工作进行评价，没有办法进行预测分析，在实际应用中存在不足。

本课题力图总结目前的科学影响力评价方法，比较各种方法的优点和不足之处，然后以微生物领域为研究对象，搜集相关数据，结合数学模型，借助于R语言，对评价结果进行比较，并对未来的发展进行预测。利用这一套思路，探究科学影响力评价的方法，最终希望能够总结归纳出一套科学影响力评价的数学模型和方法，并以参数的形式反映出来，得到评价指标，为其他相关的评价提供借鉴。

目前，在文献计量学领域，自从Hirsch提出*h*指数以后，随后又有很多学者提出许多*h*衍生指数，例如*g*指数、*R*指数等。虽然如此，目前对相关指数做预测分析的仍然不多，本研究工作既是在基于前人工作的基础上，对这些指数做一些统计分析，并构建有实用价值的预测模型，并在人才评价指标方面做一些探究。

# 1.2科学影响力及其评价方法

“影响力”的解释，一般认为是用一种别人乐于接受的方式，改变他人的思想和行动的能力。但是在国际上，对于“科学影响力（Scientific Impact）”一词并没有大家公认的定义，从“影响力”一词的解释出发，我们可以将“科学影响力”理解为，在某个学科领域的学者，对科学研究内容有了新的观点或者发现，并通过某种方式让别人了解，例如做演讲、报告或者发表文献等，这种观点被人所接受或者反驳，进而或多或少的改变了别人的科学研究方法或思维。评价科学影响力问题由来以及，对科学影响力给予客观、具体、恰当的评价，也已经成为科研成果管理的重要内容，同时，利用好的评价模型和方法做出的评价结果，已经成为人才选拔和科研立项方面的重要依据。尤其是我国进入新世纪以来，科研产出越来越多，仅在微生物领域，被SCI收录的论文数量就增长了20倍（图1.）。如何正确的看待评价、科学的开展评价、合理的利用评价，在如此多的科研成果中去粗存精，并通过分析给出对学科发展的预测，已经成为一个重要课题。但是由于科学影响力评价的方法有很多种，又涉及到不同的学科领域、行业发展、地区机构等多方面因素，科学影响力的评价是一个难题。

图2.我国微生物领域被SCI收录的论文数量

除此之外，以简单的评价方法为基础，利用复杂的方法对科学影响力进行预测，在目前阶段也是一个难题，而且也没有很好的研究结果。但是对科学影响力的评价和预测又有很大的可用之处，例如，在引进人才的时候，一方面人们倾向于引用已经成名的学术界大师，但是可能由于经费有限无法引进。另一方面，如果能够通过某种方式，找出那些有发展潜力的学者并引进，也是一个不错的选择。

1927年，两位化学家在*Science*上发文提出，图书馆员可以利用引用率数据，挑选出一小部分期刊收集起来。这个想法对后来的科学评价产生了巨大的影响力，这就是基于引用的评价方法的最初来源。目前，引用分析无处不在，被普遍应用于评价文献、学者、期刊机构等方面。一般认为，引用（citations）是度量科学影响力简单又实用的方法。本文则主要以基于引用的方法来论述科学影响力评价，这些方法主要包括*h*指数（*h*-index）以及*h*系列指数，影响因子(Impact factor)和特征因子(Eigenfactor)以及其他的相关的评价指数,每种方法都有各自的特点，但是都无法全面的完善的评价科学影响力。

在引用中，最简单的指数包括：

:表示发表的总文献的数量，能一个学者的产量，但是无法度量发表文献的重要性；

：表示获得的总的被引次数，能总的影响力，但是无法区分究竟是哪些文章的影响力高；

:表示文献的平均被引频次，能给出总体的发表文献质量，但是对于高产的学者，这种平均是有偏差的；

在这些简单指数的基础上，提出了更多复杂的指数。

## 1.2.1 *h*指数

*h*指数（*h*-index）又称*h*因子（*h*-factor）,是由美国物理学家Jorge E. Hirsch在2005年提出的，可以说，这是科学影响力评价中简单却影响力最大的一个评价指数之一。*h*指数的表述为，一位科学家发表的文献中，有*n*篇被引用的次数至少为*n*，那么该科学家的*h*指数即为*n*（图2.）。它是一个混合量化指标，既能反应一个研究者的发文数量，同时又可反应文献的质量。例如，在WoS数据库中，霍金(S. W. Hawking)的*h*指数为75，也就是说，霍金有75篇文献的被引用次数至少为75，一方面，说明，霍金至少发表了75篇文献，另一方面，他发表的文献至少被引用了752=5625次，而事实上，霍金共发表了149篇文献，总被引达到31118次，可见用*h*指数评价还是有不足之处的。

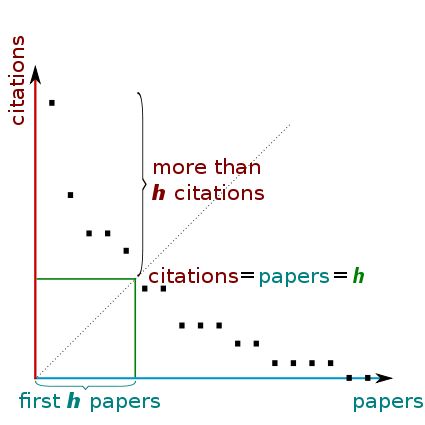


图2. *h*指数的图形描述

在Hirsch的文献中，利用数学方法他还提出了许多和*h*指数相关的参数，例如：

(1)

其中为发表的所有文献的总被引次数，*a*为比例常数，经验上讲*a*的范围介于3到5之间。

表1. 不同科学家的*h*指数及*a*值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 作者 | 研究领域 | 发文量 | 总被引 | *h*指数 | *a*值 |
| S. W. Hawking | 物理学 | 149 | 31118 | 75 | 5.53 |
| 1. J. Heeger | 物理化学 | 1239 | 101432 | 146 | 4.76 |
| P. G. deGennes | 物理学 | 392 | 45379 | 93 | 5.24 |
| J. D. Watson | 生物学 | 196 | 17804 | 50 | 7.12 |
| S. H. Synder | 生物学 | 1216 | 167762 | 211 | 3.77 |
| R. C. Gallo | 生物学 | 1150 | 106652 | 160 | 4.17 |

并且，*a*可以有更具体的解释,假设*h*篇文献中 ，篇文献的被引次数介于和之间，那么若将这篇文献的被引次数看做等差的，且 ,则总被频次为

(2)

表示贡献h指数的文献的累积被引频次；令 ,且有,则,那么比较（1）和（2）即可得到：

(3)

这个推导中，还近似地将文献的分布看成以首项是,公比是的等比级数。而, 通过这个推导，可以帮助我们由*h*指数推算出总的被引次数。

此外，Hirsch基于假设，h随时间近似线性增长，在最简单的模型中，假设一个研究者每年发表文献数量为p，随后每篇文献每年获得c次新的引用，那么n+1年后，获得的总的被引次数

(4)

假设到第y年之前发表的文献都对h指数做出了贡献，则有

(5a)

(5b)

那么可以得到

(6)

(7)

与(1)相比较，可以得出*a*与的关系。既是说，a与发表的文献量以及每篇文献每年得到的引用次数相关。

*h*指数被提出之后，立即引起了情报学界的关注，在近十年的时间里，有很多国内外学者利用不同的数据对*h*指数进行了验证。2006年，Grad Sad利用作者层面以及期刊层面的数据验证，*h*指数具有很高的启发价值（heuristic value），结果很精准，而且，需要的信息量很少。他选择55位作者在1989—2005年间在*Journal of Consumer Research*上至少发表5篇论文，计算其*h*指数，发现*h*指数和总的被引频次之间的相关系数高达0.87。在期刊层面*h*指数和影响因子的相关性为0.64。Anthony F. J.以研究组为对象，分析了荷兰147所大学的化学研究组，研究其在1991—2000年期间发表的文献的*h*指数，并发现了*h*指数和同行评议具有很高的相关性。丁楠等人则基于CSSCI 2001—2005的引文数据，统计分析了图书情报领域部分专家学者的论文和著作的被引情况，计算其*h*因子，并认为*h*指数和传统的引用评价关系十分密切，且相关系数很大。

从总体上来看，*h*指数能一定程度上反应一个科学家在某个领域的产量及影响力，计算简单，且易于理解，同时，只是发表论文的数量增加，并不产生直接影响，而且，很难通过自引的方式来增加*h*指数。但是也存在显著的局限性：

首先*h*指数只会增加不会减少，这就意味着，一个科学家的学术年龄越长，那么对应的*h*指数就会相对较大，因此不同的学术生涯之间的比较意义不大，因此h指数更适合对科学家的终身学术影响力评价，同时，这样的一种机制也使h指数无法出科学家近期的研究成果；其次，*h*指数评价的对象时一个科学家发表的所有的论文集，对于发表的那些超高被引的文献，却不出来，例如，希格斯( Peter Higgs)的*h*指数为12，但是却有3篇论文的被引次数达到2000次以上。因此，对于发表文献少却做出突出贡献的科学家，*h*指数也不适用；此外，不同的学科之间*h*指数也不具有可比性，例如在数学领域*h*指数普遍比生物学领域的低很多；最后，对于那些被引次数小于*h*的文献，*h*指数无法描述出来，而那些高被引的文献一旦被算入h指数中以后，也就无法在影响到*h*指数，因此，只有那些引用次数在*h*附近的才能影响到*h*指数。

为了解决*h*指数存在的不足，人们就尝试提出了众多*h*指数的变体。

## 1.2.2 *h*系列指数

*h*系列指数是指在*h*指数的基础上衍生出来的指数，作为h指数的完善和补充，目前主要包括*g*指数、*A*指数和*AR*指数、*hg*指数、*f*指数等。

*g*指数(*g*-index)是*h*指数的变体，是由Leo Egghe在2006年提出的，他的定义为，将某科学家发表的论文按被引次数从多到少排序，前*g*篇论文获得引用次数累积不少于*g*2次，即是说，第(*g*+1)篇论文前的累积引用次数小于(*g+*1)2，那么这位科学家的*g*指数为*g*。从定义中我们可以看出，对于同一个科学家，他的*g*指数不会小于*h*指数的值。与*h*指数一样，*g*指数越大，则说明该学者的学术影响力越大。 例如，R. C. Gallo的*h*指数为160，*g*指数为302。

相比于*h*指数，*g*指数的优点在于，那些高被引的文献引用量的增加，仍然会使*g*值增加，这样对于那些即使发文量比较少，但是却做出突出贡献的科学家，更能反映出他们的成就，例如，希格斯的*h*指数为12，*g*指数却达到99。但是*g*指数也存在一个问题，就是希格斯发表的文献远少于99篇，这样，如果按g指数的计算方法，则得到的*g*值会大于发表的论文数量，无法准确反映科学家的学术产出量。如果我们给出作为一个参考值，不难发现，当这个值很大时，就证明这个科学家有几篇文献的被引频次非常高，这可以让我们用来辨别在某个领域中发文量少但是却有突出贡献的学者。然而，这也成为了*g*指数的一个缺点，超高被引文献对结果的影响很大。于是便有人提出了*hg*指数。

*hg*指数是由S. Alonso基于*h*指数和*g*指数的基础上提出的，力图结合两者的优点，同时减小其中的不足。*hg*指数是*h*指数和*g*指数的几何平均数，数学公式为：

(8)

由于*，*所以可以很容易证明。 相对于*h*指数，如果*hg*指数有类似于*g*指数的优点，例如，当某一篇高被引文献的被引继续增加，*hg*指数仍然会增加。但与此同时，*hg*指数又不像*g*指数那样，*hg*指数可以削弱超高被引文献对结果带来的显著影响，在超高被引文献和多数高被引文献之间达到平衡。

针对于*h*指数的缺乏灵敏度、缺乏区分度和只升不降的缺陷，中国科学院国家科学图书馆的金碧辉与鲁索（Rousseau）提出了*R*指数与*AR*指数。*R*指数是指*h*指数划定的绩效核内总被引频次的平方根，其指数的数学公式如下：

(9)

其中*cj*表示的绩效内第*j*篇论文的被引频次。由于，所以。可以看出，*R*指数是依赖于*h*指数存在的，将*R*指数和*h*指数同时使用，可以弥补*h*指数的区分度不大的问题。同时，从数学上看，*R*指数和*g*指数更接近，通过比较我们还可以得出对于同一个学者的结果，同样，对于那些发文量少却又极其优秀的科学家，*R*指数也能给予很好的区分，例如希格斯的*R*指数也达到了99。

*AR*指数指的是*h*指数内每篇论文的年均被引频次总和的平方根，其数学描述如下：

(10)

其中a*j*为论文*j*的发表年龄，*AR*指数中加入了论文发表年龄这个变量，解决了*h*指数只升不降的问题。

同时，金碧辉提出，在比较一批科学家时，应该对*h*指数和*R*指数配伍，进行加权标准化

(11)

标准化之后进行比较的结果则更合理。

一般认为，*h*指数更适合评价个人产出，但是必须注意到，在不同的学科之间及某一领域内不同研究方向间可能存在差异。例如，相比于物理学和数学领域，生物学领域的顶尖学者更容易得到较高的引用，而在工程领域的顶尖学者被引次数则比较少。Hirsch认为，不同学科间*h*指数存在较大差异，这种差异一定程度上来源于该领域内文献的篇均引文数、科学家发表的文献量以及该领域内科学家的数量。因此，为了解决不同学科之间的差异，叶鹰提出了*f*指数，其数学描述为：

(12)

其中，*P, C, HCP*分别表示论文、引文和高引论文数，*TFC*是相关学科的总引文数，*CPP=C/P*表示平均引文数。由于*f*指数与学科相关，综合了质量和数量因素，因此，在不同的领域之间的比较具有参考价值。

随着研究的逐渐深入，*h*指数也被拓展应用到了期刊水平的学术评价。此时问题就凸显出来了，不同的期刊发行规模差别较大，因此，进行期刊评价时，必须考虑期刊发行周期、发文数量等因素。针对这个问题，Barendse提出了*SR*指数（Strike Rate Index）其数学描述为：

(13)

其中*h*为*h*指数，*N*为发表的论文总数。Barendse通过对农业、凝聚物理学、遗传学和数学物理学的比较，发现一个期刊的发文量越多，其拥有*h*指数越大，两者存在对数线性关系。而*SR*指数在4个不同领域中，却有相同的分布趋势和相似的质量阈值。因此*SR*指数可以在不同的期刊不同领域之间进行比较。

除了上面介绍的一些评价指数之外，还有一些其他的*h*指数相关的评价指数，这里就不详细叙述。不同的评价指数可以从不同的方面评价科学影响力，但是单一的评价指标也都有各自的局限性，往往需要把多个指标结合起来才能给出全面的评价。同时，需要指出的是，这些在*h*指数基础上衍生出来的众多评价指数，就目前来看，虽然他们能一定程度上弥补*h*指数的不足，却并没有哪一种方法比h指数好很多，所以目前，由于计算简单，*h*指数仍然被广泛使用。

## 1.2.3影响因子

影响因子（Impact Factor）在期刊评价方面是目前大家公认的而且应用最广泛的评价指标。最早在1955年Eugene Garfield就提出了影响因子的想法，在1961年开始发布Science Citation Index(SCI)之后，Irving H. Sher和Garfield一起，正式提出了期刊的影响因子。自1975年起，每年都会发布基于影响因子的期刊引证报告。我通常所说的影响因子指的是2年的影响因子，其计算方法为：

(14)

也即是说，第*y*年的影响因子为该期刊前两年发表的文章在第*y*年中被引用的次数除以前两年发表的文章总数。类似的，还有5年影响因子和10年影响因子，例如某些期刊发表的文献更倾向于有长期的被引频率，因此其5年的影响因子就会偏大。

影响因子由于其计算方式简单，含义清晰明确，所以是目前最为广泛使用的评价期刊科学影响力的方法。影响因子其实是篇均被引频次，因此，该指标可以从整体上体现某个期刊的发文水平，正确反期刊的质量。除此之外，影响因子是一个随时间动态变化的指数，只与近两年内的发文质量有关，所以能够很及时的反映期刊质量的变化。

但是，影响因子简单的计算方法，也带来了诸多局限性:首先，它反映的是期刊的整体水平，而无法描述期刊中具体文章的科学影响力，例如，相对低水平的文章发表在高水平的期刊上，就很容易高估这篇文献作者的水平，众所周知的黄禹锡造假事件让学术界一片哗然。同样，它也无法反映期刊同一期的所有论文之间的差异,同一个期刊中，往往是一少部分高质量文献贡献了绝大部分被引次数，早在1997年，Seglen的研究就指出，将期刊中的文献按由多到少的被引次数排序，前5%的文献贡献了约20%的被引次数，前15%贡献了50%的被引次数，而前50%的文献贡献了接近90%的被引次数。由于影响因子是基于引用计算的，因此存在人为操纵的可能，例如发表更多的综述以换取更多的引用次数，或者增加自引次数，这些都可以提高影响因子，*Nature*就曾经报导，巴西的几家科学期刊合谋，利用相互引用论文的方式，快速提高影响因子；另外，在不同的学科之间以及不同的领域之间，同样无法相互比较影响因子；除此之外，对于那些文章偏向于长期被引用的期刊，影响因子也无法正确地反映期刊的长期影响力；在计算的同时，影响因子将所有的引文看作是等同的，这一点当然不够合理，来自顶尖期刊的应用和来自普通期刊的引用，自然是不一样的；虽然说是影响因子，但是他反映的只是直接引用，而没有考虑间接引用因素，因此也无法全面的描述期刊的科学影响力。

影响因子的诸多缺陷，在学术界引起了很多讨论，Seglen就认为用影响因子来评价科学研究很不合理，因为只有在绝对引用平均的情况下它才正确，这显示和实际存在很大差距，因此他认为用影响因子评价科学研究是没有代表性的，因此也就无法利用影响因子评价某个科学家的科学影响力。Opthof认为，用影响因子评价单独的文章、科学家以及研究部门，是缺乏科学基础的。

1.2.4 特征因子

针对于影响因子的缺陷，就有人提出了特征因子(Eigenfactor)。2007年，Bergstorm提出了特征因子的概念，其基本假设是：某个期刊被其他高影响力期刊引用的次数越多，则其影响力也越高。影响因子在计算期刊的引用次数时，将来自所有刊物的应用都看作是一样的，但实际上，来自于诸如*Nature*和*Science*这样的期刊的引用应该具有更大的分量。Bergstorm和West从这个思路出发，考虑到不同层次期刊的引用权重，在除去自引后，通过引文构建引用网络，以类似于Google所用的网页排序和排名的PageRank算法，迭代计算出期刊的权重影响值，实现了引文数量与价值评估的综合评价。

特征因子的原理是这样的：假设一位读者随机选取某年（比如2010年）出版的一份刊物（刊物1），在其中任选一篇文章，在其参考文献中，任意选择一篇。跟随这篇参考文献，进入到下一份期刊（刊物2），然后再从2010年出版的刊物2中，任意选择一篇文章，然后继续随机选取引文，进入刊物3。以此方法，不停地重复下去。显然，如果期刊的影响力越大，被更多的文献引用，则读者进入这个期刊的次数越多。因此，定义读者进入某个期刊的几率的百分值，就是该期刊特征因子的分值。例如，*Nature* 2012年的特征因子值为1.5654，也即是说，按照上面的方式，一位读者有1.5654%的概率进入*Nature*。

特征因子解决了影响因子的一些局限性。在计算中，每一次施引的重要性为施引期刊的影响力除以其施引总数，这种算法在不同的学科中都有通用性，因此，在跨学科之间以及不同的文献类型之间，特征因子仍然具有可比性。例如，来自一篇具有大量引文的综述的引用，其重要性要比来自一篇只引用相关文献的文章的引用重要性要低。同样，我们也可以看出，特征因子评价的是期刊的整体影响力，会受到期刊刊文数量的影响。对于两种文献质量相当的期刊，发文量大的一方，就会具有更高的特质因子分值。由于计算方法的不一样，影响因子和特征因子的排序会产生很大的差异。例如，在 Journal Citation Report数据库中，2012年，*CA: a Cancer Journal for Clinicians*的影响因子为153.46高居第一，但是特征因子只有0.05136，排名第319，而*Proceedings of the National Academy of the Sciences of the United States of America（PNAS）*的影响因子为9.74，排名168，但是特征因子高达1.5462，仅次于*Nature*排名第二。因此，在对期刊的科学影响力评价时，采用的方法不同，很可能得到的结果差别很大。通过比较可以发现，影响因子和特征因子的差异，主要是由于发文量差异造成的，例如，*PLOS ONE*杂志2012年共发表了23406篇文章，其影响因子很低，仅为3.73，但是特征因子却高达0.7784排在第6位，相比之下，*CA: a Cancer Journal for Clinicians*虽然影响因子高达153.46，但是2012年一共发表25篇文章，这就导致其特征因子十分低。为了消除发文量带来的影响，论文影响分值（Article Influence Score）的概念被提出来了，它的算法为特征因子除以期刊所发表论文标准化值（以所有期刊的论文总数为1），其平均值为1，若分值高于1，则说明文章的影响力高于平均水平，反之则反。

通过上面的比较，我们可以发现，影响因子更适合用来评价期刊中文献的平均质量，而特征因子则更适合评价一个期刊在某个领域中的整体影响力。但是，如果我们仅仅用论文期刊的影响因子或者特征因子去评价某位科学家时，评价结果的可靠性是值得商榷的。

# 1.3科学影响力的预测方法

上文中，我们对科学影响力评价中一些有影响力的方法进行了比较，不同的方法都有各自的优点与不足，因此一个完善全面的评价结果，往往要将多种评价方法结合使用。即便是能给出一个较为全面的评价结果，这些方法只是基于已有的数据，对现状进行评价分析，而无法对以后的发展情况做出预测。若想解决这个问题，则需要利用复杂的数学模型，以这些简单的评价指数为基础，利用统计分析的方法进行预测，这也是现阶段我们想解决的问题。目前，这方面的研究仍然比较少，我们以两个例子为基础做详细的介绍。

## 1.3.1 线性回归分析法

线性回归是利用线性回归方程中的最小二乘函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种方法建立的模型简单，但是结果却可以很可靠。Konrad P. Körding等人，利用线性回归方法，对h因子做出预测得出了可靠的结果。

线性回归分析理论模型如下，假设有一系列自变量，同时，对应的因变量和他们的相关系数分别为，于是线性回归方程就可以写作

(15)

其中表示误差，它服从均值为1，标准差为的正态分布，的值和时间有关。当变量*x*的数量很多时，我们可以假设其中有一部分是共线性的，于是可以将变量合并，以减少回归分析的复杂性。这一步可以通过弹性网络正则化来完成，即是让下面公式取最小值：

(16)

其中，,用来表示模型的复杂性（值越大，模型越简单），表示共线性变量的程度（接近于0代表很多共线性，接近于1则表示没有相关性）。在Körding的研究中，他们另，然后找到合适的值。

Körding的研究中，他们从h指数大于4的研究者中，通过数据筛选，选出了3085名神经学科学家，57名果蝇研究者和151名进化学家，统计了他们的文献发表相关的数据，然后利用上面的思路，通过回归分析和机器学习法，对h指数进行了预测。在对神经学家的*h*指数预测时，得到如下预测结果

1. 预测下一年的*h*指数（）:
2. 预测5年后的*h*指数（）：
3. 预测10年后的*h*指数（）：

其中*n*表示发表的所有文献的数目；*h*为当前的*h*指数，*y*表示从发表第一篇文章开始到现在的年数；*j*表示在多少不同的期刊上发表过文章；*q*表示在*Nature, Science, Neuroscience, PNAS，*以及*Neuron*上发表的文章的数量。

从的值可以看出，在对下一年预测时，能得到相对准确的值，随着时间的变长，则显著变小。同时，他们还发现，在对果蝇研究者和进化学家的h指数预测时，的值要小很多。但是这样的结果也比只利用*h*指数进行预测要好很多。而且我们可以看出，随着时间的增加，*y, j, q*的相关系数在增加，而h的相关系数却在减小。

尽管我们难以从这个模型中得出很有说服力的结论，但是，却给我们带来了一些启示。对*h*指数影响最明显的还是发文量*n*，另外，时间越久，在顶尖刊物上发表的文章对*h*指数的影响越大。相比于利用单一的指标评价，Körding的研究中，引入了更多的变量，如果想操控评价结果将会变得很困难。

## 1.3.2基于引用动力学模型的预测

上面的研究工作，一定程度上仍然是以*h*指数为基础的，所以仍然具有*h*指数的局限性。为了完善预测方法，Albert-László Barabási等人则深入研究了引用动力学并根据其机制，建立了数学模型。他们的研究认为，所有的文章，不论是来自不同的期刊和学科，他们的被引用的方式都是一样的。这一发现有助于我们了解文献引用的机制以及科学影响力是如何发挥作用的。

Barabási的研究首先搜集了*Physical Review (PR)*杂志从1893年到2010年发表的全部463348篇文献，并统计了每一篇文献的被引情况随时间变化的关系。他们发现，虽然绝大部分文献很少被引用，但是确实有少部分被引频次非常高。在每一篇文献的被引变化历程中，他们发现了3个基本的机制：

①高被引文献更容易被再次引用，因此一篇文献第*i*次被引的概率和它之前的总被引频次成正比；

②一篇文献的新颖性会随着时间的慢慢衰退，其长期的衰退情况可以通过对数正态分布生存概率描述：

（I）

其中*t*表示时间，即文献发表了多久；是即时性(immediacy)，表示文献被引概率最大时的时间；则表示文献的寿命，与衰亡率有关。

③如果用，适合度(fitness)来表示一篇文献固有的属性，用来描述文献的新颖性和发现的重要性，如果我们将适合度看作一个集体量(collective measure)表示公众对一篇文章的反映，那么就可以得到简化模型，一篇文献*i*在发表后*t*时刻被引用的概率为：

(II)

基于这3个假设，我们对引用的分布有了详细的数学描述，通过求解之后可以得到文献*i*在*t*时刻的累积被引频次为：

（III）

其中为累积标准正态分布。*m*代表每篇新文献的平均参考文献数量，用来描述总的发文量的增长率，*A*是标准化的常数。因此我们可以看出, *m*, , *A*都是通用的参数，对于所有的文献都是一样的。该等式描述了一个最小引用的模型，而且已经将3个基本机制全部包括在内。从这里面我们可以看出，一篇文献的被引历程由3个基本参数描述：相对适合度, 表示一篇文献相对其他文献的重要性；和。如果用, , 那么我们可以得到：

(IV)

这个等式表明，一篇文献根据其特有的修正之后，它的被引用历程都是符合累积正态分布曲线的。因此如果知道一个文献的被引用历程，我们就可以求出其特有的参数。

如果我们令等式III中的，就可以得到一篇文献最终被引用的频次：

(V)

从这里我们可以看出，一篇文献最终的被引用次数只和相对适合度有关，也就是说，一篇文献相对于其他文献重要性越高，创新性越大，则最终得到的引用次数也会越多。

以上面的工作为基础，Barabási等人借用天气预测和数据挖掘中的一些方法，将一篇文献中的某一段时间内的引用历程用作训练集，来计算出，然后利用等式III来预测其，利用等式V来预测其，并根据噪音给出预测区间，他们得到了很好的预测结果，而且如果将训练集的时间延长，结果的准确率也会更高。经过统计他们发现，这个模型对文献25年后的被引情况的预测准确率高达93.5%。

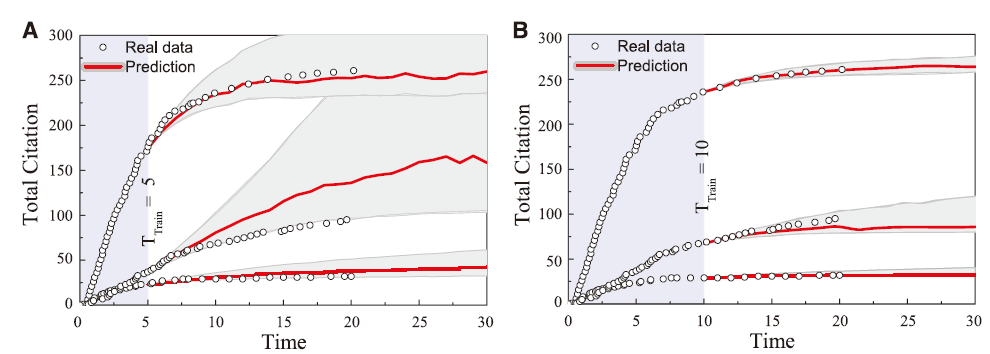


图3. Barabási等人得到的预测结果。A图的训练集时间为5年，B图为10年（左边阴影区），白色点为实际数据，实线表示预测的最有可能的结果，灰色区域为预测区间。

当然，Barabási也指出，这个模型是有局限性的，对于那些过了一段时间之后又再次被大量引用的“Sleeping Beauty”这样的文章，这种模型是无法预测出来的。不论怎么说，他们深入引用动力学机制的研究，能让我们对长期的科学影响力有更好的了解。

就目前的研究现状来看，除了Garfield提出的影响因子之外，其它的一些评价方法都是近十年内提出的，可以说，科学影响力的评价仍然处于起步阶段，而预测分析则刚刚是萌芽阶段。影响因子和*h*指数在评价中，仍然被普遍使用，一方面是由于二者的计算相对简单，另一方面是由于人们已经习惯使用这两种方法，而新诞生的方法又没有特别大的优势，所以在普遍使用中，仍然无法完全替代这两种方法。归根结底，新的研究方法并不是因为太复杂而不能被接受，而是突破太小，结果也不那么可靠。就比如我们需要修建核电站，不能因为相对论太复杂，人们看不懂，而去使用简单的牛顿力学原理，其实我们只需要告诉别人，我们的结果是绝对可靠的，你只需要看结果就行了。正是因为现在缺乏这样的可靠性，才会给评价工作带来了困扰。

从本质上来看，*h*指数、影响因子和特征因子分别出自不同的体系，具体涵义也不同，因此在评价中相互独立不可取代。但是*h*指数和*h*系列指数则有很大的重叠性，是可以互相替代的。在理论上，*h*系列指数也都是从*h*指数衍生而来，因此也并不能完全摆脱*h*指数的局限性。从这个角度来看，虽然出现了多种h系列指数，但是并没有能推进科学影响力评价方法的创新，因此，仍然需要探索新的评价方法。

相比于评价方法，我认为预测是一个更值得探索的问题。评价其实可以看做是对已有结果进行描述，可是预测则必须要深入理论研究，只有分析清楚内在的机制，才能做出好的预测。无疑，对科学影响力的预测分析能将研究工作推入一个新的深度。Barabási的研究工作为我们提供了很好的思路，他提出的引用动力学机制具有普遍的适用性，因此，可以作为后续研究工作的基础。我们期待着，深入的研究不仅能将科学影响力的评价推向一个新的平台，也能为未来的发展做出很好的预测。

# 1.4研究方法与技术路线

## 1.4.1 研究方法

本课题主要分为两部分，第一部分主要在比较各种科学影响力评价方法的不同，选出适合建立数学评价和预测模型的参数。第二部分则主要以微生物领域的数据为基础，比较我国和其他国家的异同，整理出评价思路，并根据实际情况，优化模型，给出更好的评价和预测方法。具体研究方法包括：

1. 文献计量法：搜集微生物领域相关数据，得到h指数和h系列指数、影响因子、特征因子等评价指数，通过文献计量方法，从人才、期刊、机构等角度描述我国微生物领域的发展现状。
2. 利用R语言，对数据进行处理，建立微生物领域文献引用相关数据库。
3. 比较分析法：通过对比，得到我国微生物领域与其他国家存在的差异，同时，对比各种评价方法，分析出各自的优缺点，以此为基础建立更好的评价思路。
4. 统计学方法：用搜集的数据，通过回归分析构建模型，对评价指标进行建模，并给出预测分析。同时，利用Barabási的工作中提出的模型，借助于机器学习的方法，对微生物领域的数据进行预测分析。

## 1.4.2 技术路线

1. 首先，采用文献调研的方法，以微生物领域为对象制定检索策略，利用WoS数据库，检索文献，并将文献及引用相关数据下载下来，并根据作者、机构、期刊等角度分类整理。
2. 利用Excel和R语言等，将数据归类，并按格式构建简单的数据库，以便下一步分析。
3. 利用统计学方法，借助R语言，用最小二乘法对h指数进行回归分析，并根据机器学习的方法，作出预测分析。
4. 先用微生物领域的数据对引用动力学模型进行验证，分析引用数据，并探究能否进行优化，分析结果，并利用机器学习的方法进行预测，从人才、机构的角度预测科学影响力的发展趋势。
5. 在以上工作的基础上，借助比较分析法，归纳总结出一套评价科学影响力的方法和思路。

通过对微生物领域相关的科学影响力进行评价，比较我国和其他国家发展的异同。利用数据对预测模型进行验证，并尝试对模型进行优化，反过来再次比较预测结果的差异。以此为基础，归纳出一套评价和预测科学影响力的方法思路。

# 1.5关于本论文

## 1.5.1本论文的目的

## 1.5.2论文的内容安排

第二章 数据库的构建与描述

# 2.1数据来源检索策略

不同的学科，引文的分布可能有很大差别。本论文以微生物领域的文献为研究对象，因此，首先选取了一些微生物领域的学者的文献数据。这些人员的选取，依据FACULTY1000网站中，分类为“Microbiology”下面的部分研究人员[[1]](#footnote-1)。然后，在Web of Science网站中，分别检索这些学者的发文及引文数据，精炼类型为“Article”和“Review”，在“创建引文报告”这一功能中，将所有的文献及详细的引文数据保存为Excel的格式下载下来，这部分共搜集了212名研究者，共22514篇文献的数据。此外，还需要其它学科的数据作为对照分析。考虑到微生物学的交叉性比较强，为了尽量减少研究人员有重复，最后选择F1000网站中“Neuroscience”分类下面的研究人员的发文数据，这部分共327人，33155篇文献，样本数据和对照数据中没有重复。

# 2.2数据库的结构

在原始数据中，保留了很多字段，例如文献的标题，作者，出版年等信息，而我们分析的时候，只选用标题、作者、来源出版物名称、出版年、合计应用次数、每年的平均被引次数以及从1950年到2014年每一篇文献详细的被引频次这些字段，做后续的分析工作。经过简单的处理，我们在原始数据的基础上，构建了基础数据库。基础数据库格式如图，对应的分类字段如表。



图. 原始数据格式

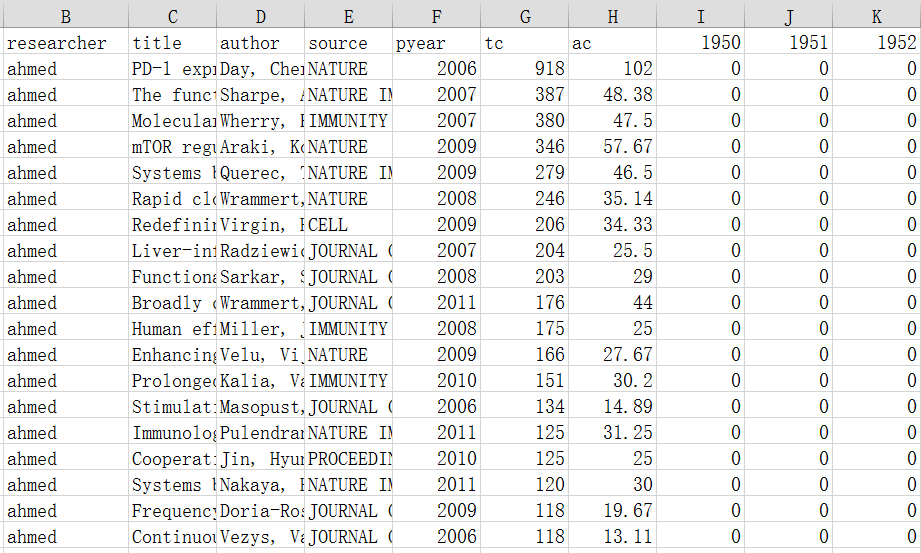


图.基础数据库结构

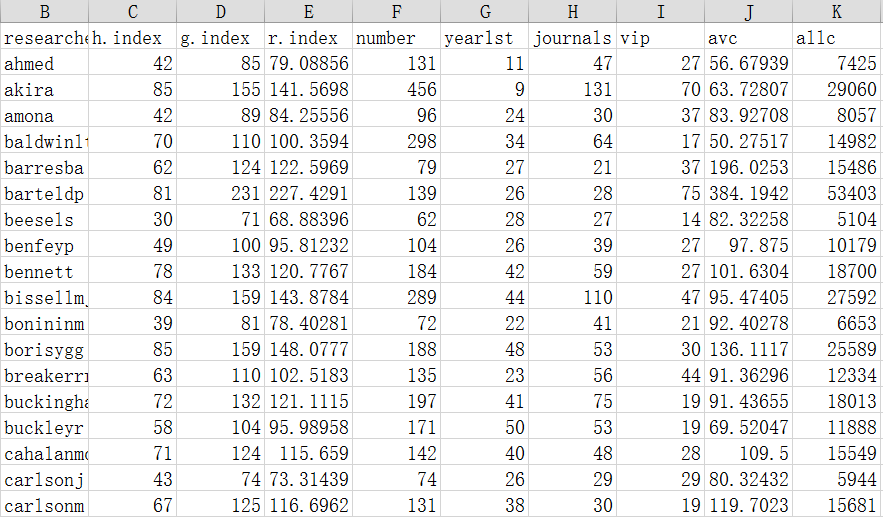
表1. 分类字段极其对应的内容

|  |  |
| --- | --- |
| 分类字段 | 对应内容 |
| researcher | 研究的对象，即检索时的作者 |
| title | 检索到的文献的题目 |
| author | 一篇文献中署名的所有作者 |
| source | 文献的来源，即其发表在哪个期刊上 |
| pyear | 文献发表的年份 |
| tc | 每一篇文献的总被引频次 |
| 1980-2014 | 对应的年份中，每篇文献每年的被引频次 |

在基础数据库的基础上，需要计算出每个作者所对应的各种指标的情况，这里主要选取了如表2中的若干指标，再此基础上，构建出了二级数据库，格式如图

表2.二级数据库分类字段所对应的内容

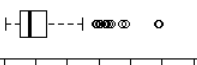
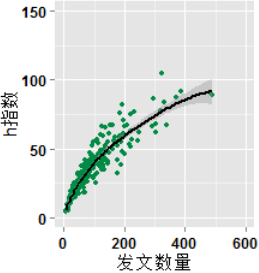
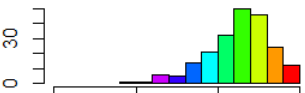
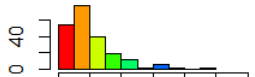
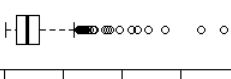
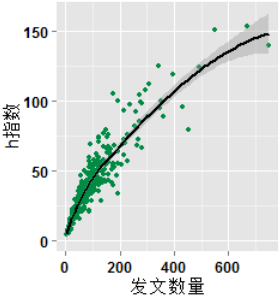
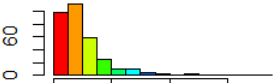
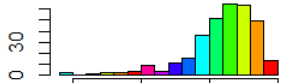
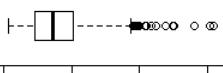
|  |  |
| --- | --- |
| 分类字段 | 对应内容 |
| researcher | 研究的对象，即检索时的作者 |
| h.index | h-指数 |
| g.index | g-指数 |
| r.index | R-指数 |
| number | 发表的文献数量 |
| years1st | 距离发表第一篇文献的时间（单位：年） |
| journals | 一共在多少个期刊上发表过文章 |
| VIP | （very important paper）指在*Science, Nature, Cell*等高影响力上发表的文章的数量 |
| allc | 所有文献总被引频次 |



图，二级数据库格式

# 2.3 数据库的描述

在具体的分析之前，我们先对数据库的各种指标进行了统计分析，从而对数据有一个整体的了解。首先分析了在样本数据和对照数据中，每个研究者的发文数量及h指数关系图，如下图



*a*

*b*

图#. 发文数量及*h*指数关系图。*a*图为样本数据，*b*图为对照数据

从图中可以看出，随着发文数量的增加，h指数增大，且相关性十分明显，这也可以从后面的相关关系图中看出来。虽然样本数据和对比数据学科不同且数据量也不同，但是h指数与发文数量的总体分布是一致的，不同点在于，当发文量较大时，神经科学领域的研究者能得到更大的h指数。例如，当发文量大于200时，微生物学领域的学者平均h指数为71.08，而神经科学平均h指数为93.63，差异十分明显，如表#。引起该差异的可能原因是，在发文数量上，虽然神经学领域有部分研究者发文数量高达400以上，但是两个学科总体的发文数量平均值相当接近，微生物学为106.20篇，神经学为101.39篇。

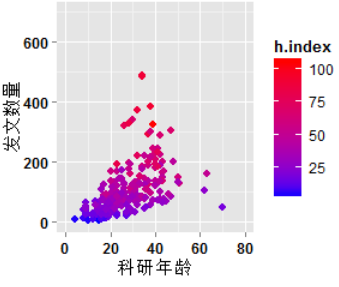
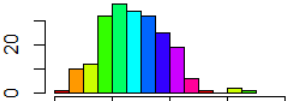
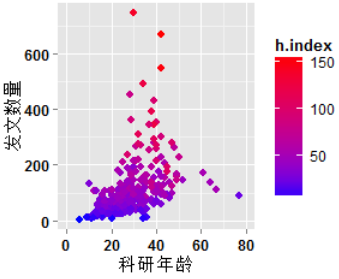
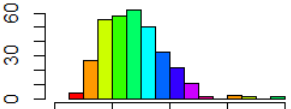
表#

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 发文数量区间 | h指数平均值 | | 数据数量 | | 方差比 | 自由度 | p值 |
| 样本 | 对照 | 样本 | 对照 |
| 总体 | 37.76 | 40.98 | 212 | 327 | 1.70(>1.00) | 521.75 | 0.091 |
| ﹤200 | 33.51 | 35.27 | 188 | 295 | 1.27(>1.00) | 431.84 | 0.23 |
| ≥200 | 71.08 | 93.63 | 24 | 32 | 2.91(>2.43) | 51.37 | 0.00015 |

随后，又分析了科研年龄、发文数量以及h指数之间的关系。结果如图#，从图中可以看出，随着科研年龄的增长，发文数量增加，与此同时，h指数变大。发文数量--科研年龄的分布比较分散，其中还有一些偏离较多的点，如图#a图中箭头所指，该处的研究人员为*Willem de Vos*，荷兰人，在微生物学领域取得了极大成就，并于2008年获得荷兰最高科学荣誉——斯宾诺莎奖。图b中箭头所指，该处的研究人员为*Mark P. Mattson*，美国人，被认为是神经退行性疾病研究领域的领导者。从总体来看发文数量—科研年龄分布并没有因为学科不同而产生太大差异，即使对数据进行分层检验，二者的分布情况也非常接近，微生物学平均科研年龄为28.14年，神经学平均科研年龄为28.24年。

*b*

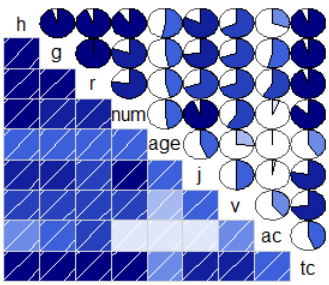
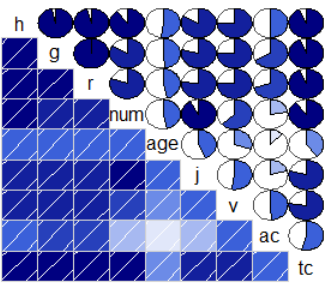
*a*



图#.科研年龄与发文数量及h指数的分布关系。a为样本数据,b为对照数据。

为了分析各个参数之间的关系，我们有做了相关性分析，结果如图#，从图中可以看出，h指数、g指数和r指数相关性十分明显，这也说明这三个指标在文献计量学描述方面具有一致性。另外，h指数、g指数与r指数和总被引频次（allc）也存在明显的相关性，但是与平均被引频次相关性则不那么强，从图中还可以看出，发表的文献越多，则分布的期刊种类越广，进而影响h指数。另外，发表在重要刊物上的文献数量，对h指数也造成较大影响。

Hirsch在提出h指数的时候，同时提到一个人的总被引频次，是*h*2的a倍，其中a是一个稳定的常数，介于3到5之间，为此，我们又分析了a值的情况。从相关性分析中（表#），可以看出，a值（hac）只与平均被引频次存在一定相关性，而与其他参数无明显相关性。从图#中可以看出，a主要集中在2.5与4.5之间，在212个样本数据中占76.9%（163个），而在2到6之间，则占了91.5%

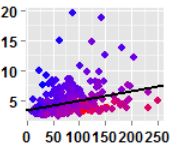
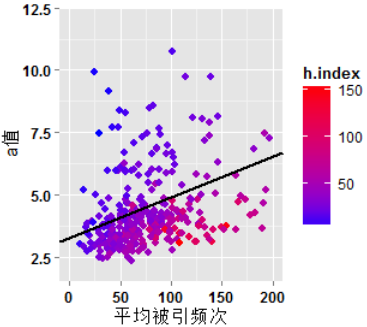
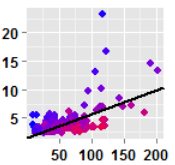
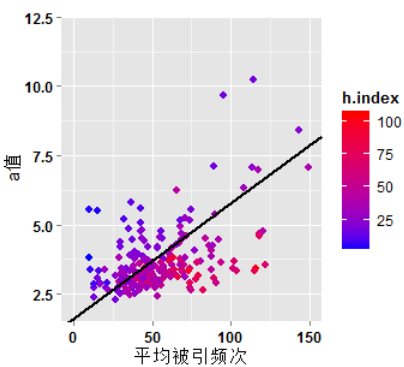
图#.各参数之间相关性分析。左侧为样本数据，右侧为对照数据

（194个）。图中样本数据拟合直线的斜率为0.042，a值随平均被引频次增加十分缓慢。在对照数据中，a值与其他参数都无较大相关性。图中可以看出，a值仍然集中在2到6之间，在327个数据组占85.0%（278个），不过与样本数据相比较，在大于5的区间内，共分布有75个数据，占总体的22.9%，而样本数据的这一比例仅为13.7%。而后，进一步分析了a的平均值。样本数据的平均值为4.06，而对照数据的平均值为4.57，经过t检验显示二者的平均值存在显著差异，p值为0.010。由此可以看出，在搜集到的数据中，相同的h指数下，神经学的研究者获得的总的引用量更大一点。

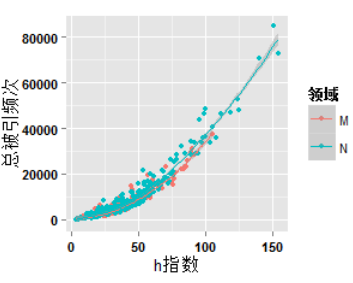
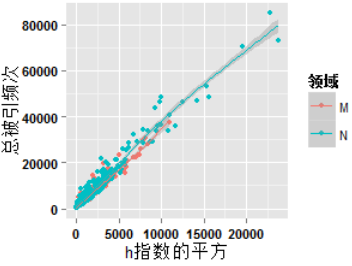
表#.a值相关性表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 相关系数 | h | g | r | num | age | j | v | ac | tc |
| 样本a值 | -0.29 | -0.094 | -0.063 | -0.26 | -0.31 | -0.27 | -0.047 | 0.58 | -0.090 |
| 对照a值 | -0.35 | -0.18 | -0.16 | -0.29 | -0.35 | -0.29 | -0.14 | 0.31 | -0.15 |

/



图#.平均被引频次与a值的分布情况

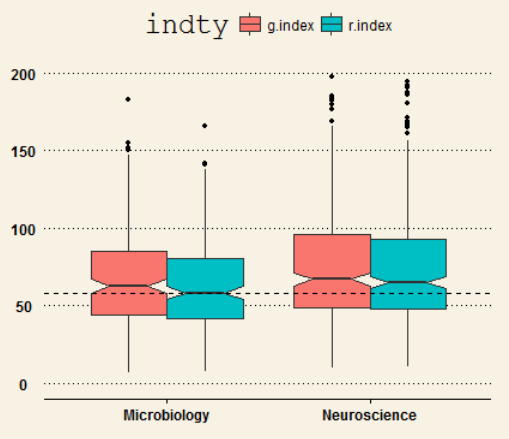
图#.h指数与总被引频次的关系。其中M代表微生物学，N代表神经生物学。

同样，当我们分别以h指数以及h指数的平方为横坐标，总被引频次为纵坐标作图的时候（图#），结果也与上面的情况相符。当横坐标为h指数的时候，该曲线近似为抛物线，当横坐标为h指数的平方的时候，得到一条直线。另外，还可以看出，当h指数相同的时候，神经生物学领域拥有较大的总被引频次。

最后，又对数据中文献的发表年份、g指数、r指数等参数进行了分析。首先从发文年份来看，样本数据和对照数据的分布一致，并没有存在显著差异。图中统计的是从1980年开始。由于前面提到，由于两个领域的研究者平均科研年龄均为28岁，所以在1980年之前发文数量都很少，其中微生物领域数据582篇，占总体的2.6%，神经学领域的数据878篇，同样占总体的2.6%。

图#.文献发文年份的分布情况。a图为样本数据，b图为对照数据。

在相关关系图中，可以看出g指数、r指数都和h指数存在很强的线性关系，这其实也是由他们的定义决定的。g指数指的是前g篇文献的总被引频次不小于（g+1)2，这就决定了g指数一定会比h指数大。而r指数指的是在h指数范围内的所有文献的总被引频次的算术平方根，通常情况下存在 h < r < g的关系，但是当作者的发文数量较少但是总被引频次特别大时，就会存在g < r的情况。必须注意到，g指数受到发文数量的限制，既是最大的g指数也不会超过总的发文数量，但是r指数则不受这个限制，因此才会出现某些情况下g指数小于r指数。从图#可以看出，两个领域的g指数在总体分布上比r指数要大，而神经生物学领域的指数又比微生物学领域要高，这也与h指数的分析结果一致。



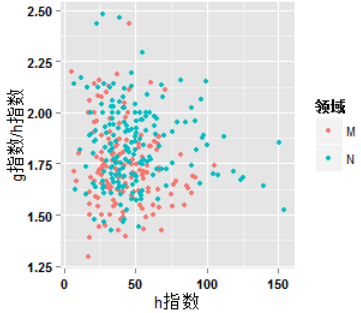
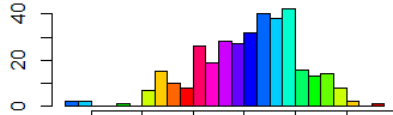
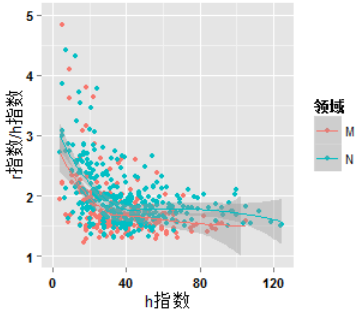
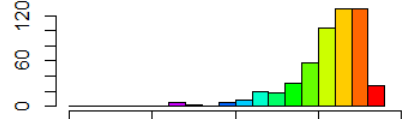
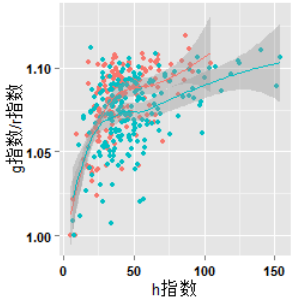
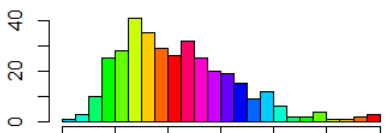
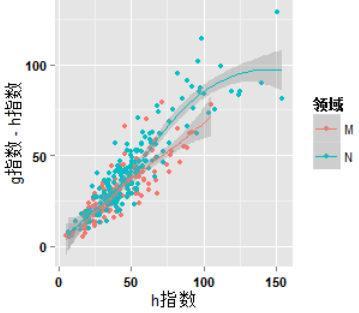
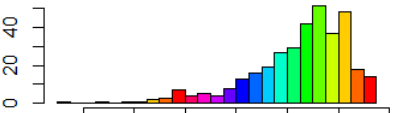
图#.g指数与r指数的分布情况

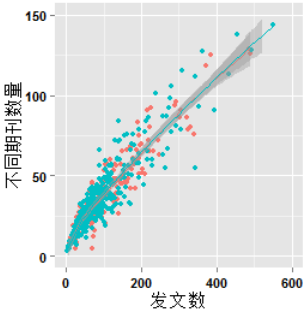
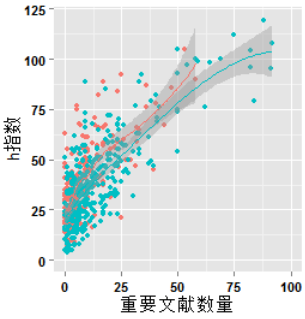
由于g指数会受到发文数量的限制，从而带来干扰，所以必须分离处那部分不受限制的数据进行单独分析。在样本数据中，g小于r的数据有51个，占总体的23.1%，而对照数据中，有139个，占总体的42.8%。从表中可以看出，在的数据中，这些研究者都拥有较少的发文数量，但是平均被引频次却很高，同时h指数却偏小，且a值很大，综合推断，这部分学者虽然发文量不多，但是却有少数文献得到了大量的引用。

表#.根据g与r的关系分组比较结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分类 | 领域 | 数量 | 平均发文数 | 平均被引 | 平均h指数 | 平均a值 |
| g ≤ r | 微生物学 | 51 | 37.98 | 81.15 | 22.65 | 6.31 |
| 神经科学 | 140 | 49.96 | 95.69 | 29.91 | 5.87 |
| g > r | 微生物学 | 161 | 127.90 | 51.05 | 42.58 | 3.34 |
| 神经科学 | 187 | 139.90 | 66.49 | 49.31 | 3.60 |

图#中进一步描述了三者之间的关系。从a和b图中可以看出，随着h指数的增大，的值并不存在明显规律，但是 随着h指数的增大而增大。而c图则表明，随着h指数的增大， 的值逐渐减小，r指数与h指数大小越接近。在d图中，我们选取的是所有g≥r的数据，可以看出随着h指数的增大值逐渐变大，但是从提来看，g指数与r指数的值仍然比较接近。



图#.h指数、g指数、r指数三者之间的关系

图#.不同的期刊数量与很重要文献数的分布情况

最后，我们对发文的不同期刊的数量以及重要文献数量的分布情况进行了简单分析。如图#a，很容易理解，发文数量越多，则发表在不同的期刊的数量也越多。发表在重要杂志上的文献，往往获得更多的引用，因此，重要文献数量越多，则h指数越大。

# 2.4 小结

本文旨在完成通过引用分析完成对科学影响力的评价，为了实现这一目的，本文首先选取了微生物学领域的部分研究者的发文数据作为研究对象，以Faculty1000网站的数据为参考，选定研究者的范围，其中微生物学的研究者作为样本数据，神经科学的研究者作为对照数据，并在Web of Science网站中检索到这些研究者的发表的文献以及详细的引文数据。这些数据构成了本课题研究的原始数据，为后续的分析打下了基础。

随后，我们借助R和Excel这两个工具，对数据进行整理分析，选取保留必要的字段，构建成初级数据库。在此基础之上，对初级数据库中的研究者、h指数、g指数、r指数、发文数量、科研年龄、发文的不同期刊数量、重要文献数量、平均被引频次以及总被引频次进行了统计分析，并将这些字段的统计结果构建成数据框，作为二级数据库。本章的后半部分，对二级数据库进行了详细的描述分析。从分析中可以看出，首先h指数、g指数和r指数三者之间存在很强的线性关系，这表明他们在文献计量学意义上具有很强的一致性。同时，当我们对数据进行不同方式的分组分析时，他们之间也体现了各自的差异。其次，可以看出，h指数主要受到发文数量、发文的期刊数量、重要文献数量以及总被引频次的影响，而科研年龄和平均被引频次对其影响相对较弱。通过对二级数据库的分析，我们对数据库中各个参数之间的关系有了更深入的了解，这为我们后面的模型分析工作提供了便利。

第三章 模型构建及科学影响力的预测

本课题的旨在利用统计学模型方法，对科学影响力进行评价和预测。我们主要选取了h指数、g指数和r指数进行线性回归分析构建模型，通过对数据分析，对回归结果进行优化，然后利用优化的模型，对上述三个指数进行预测分析，并分析预测结果的准确性。随后，我们利用引用动力学模型，进行非线性最小二乘拟合，对单篇文献的影响力进行了预测。

# 3.1 线性回归模型的构建及优化

## 3.1.1 *h*指数简单线性回归分析

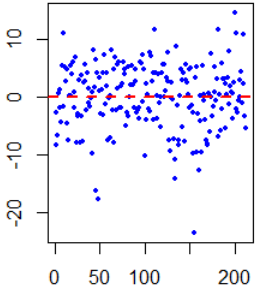
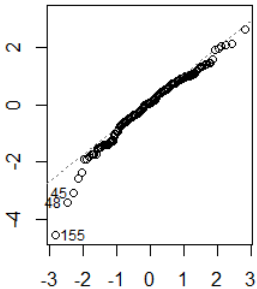
*h*指数作为最常用的指数，我们首先对其进行了模型分析。根据上一部分的相关性分析，我们知道，与h指数相关的参数主要有发文数量、科研年龄、发文的不同期刊数量、重要文献数量以及总被引频次。我们将这几个参数作为回归分析的变量，对h指数进行线性回归分析。

在微生物学数据中，我们得到了如下的回归方程：

(公式#)

拟合参数相关系数如表#，我们可以从p值看出，在我们的回归分析模型中，发文数量、科研年龄以及总被引频次与h指数有显著相关性，该模型的拟合度，具有很高的拟合度。图#显示了样本的残差分布以及Q-Qplot图。从残差分布图中可以看出，拟合模型的残差较为均匀的分布在直线的两侧，而且标准化的残差与理论分位数的分布也近似为一条直线，说明残差的分布是近似符合正态分布的。但是仍然可以看出，有部分数据偏离较大。随后，我们又分析了预测值与实际值的分布图，图中可以看出当h指数较小和较大时，预测值发生了偏离，并没始终沿直线分布，说明该模型仍然存在误差，需要进行修正。

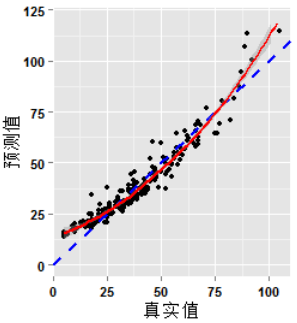
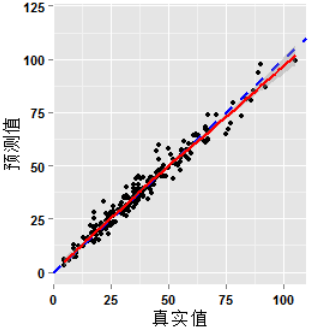
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | 11.26 | 1.22 | 9.25 |  |
|  | 发文数量 | 0.058 | 0.016 | 3.55 |  |
|  | 科研年龄 | 0.24 | 0.040 | 5.87 |  |
|  | 发文期刊数 | 0.039 | 0.047 | 0.84 | 0.40 |
|  | 重要文献数 | 0.11 | 0.056 | 2.05 | 0.042 |
|  | 总被引频次 |  |  | 11.95 |  |

标准化残差

残差

图#.样本数据拟合残差分布以及Q-Qplot图

图#.真实值与预测值的分布图。a图为原始模型，b图为修正后模型。

在第二章的数据分析中，我们发现发文数量和h指数之间，以及总被引频次与h指数之间，并不是线性关系，并且，从上述分析中，发现科研年龄对h指数的影响并不明显，基于此，我们队模型进行了修改。得到如下的修正模型：

拟合参数的相关系数如表#。从表中可以看出，修正后的模型中，科研年龄对h指数的解释并不明显，因此，考虑去掉科研年龄这一变量。

表#

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | -7.15 | 0.90 | -7.94 |  |
|  | 发文数量 | 2.42 | 0.25 | 9.84 |  |
|  | 科研年龄 | 0.031 | 0.029 | 1.07 | 0.29 |
|  | 发文期刊数 | -0.11 | 0.030 | -3.62 |  |
|  | 重要文献数 | 0.088 | 0.036 | 2.46 |  |
|  | 总被引频次 | 0.34 | 0.018 | 18.94 |  |

最后，得到的模型如下：

拟合相关参数表如下（表#）。表中可以看出，保留下来的变量对h指数的解释均有明显显著性，该模型的拟合度为，而且，从图#.b图可以看出，最后修正之后的模型，真实值与预测值更好的分布在了直线的两侧。综合看来，对于样本数据来说，该模型具有较高可信度。

表#.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | -6.91 | 0.87 | -7.92 |  |
|  | 发文数量 | 2.52 | 0.23 | 11.05 |  |
|  | 发文期刊数 | -0.11 | 0.029 | -4.01 |  |
|  | 重要文献数 | 0.084 | 0.035 | 2.37 |  |
|  | 总被引频次 | 0.34 | 0.018 | 18.94 |  |

同样，我们对神经生物学的数据进行了拟合，得到了如下的方程：

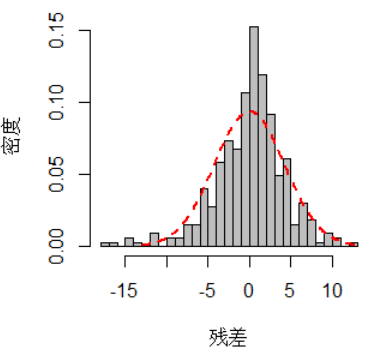
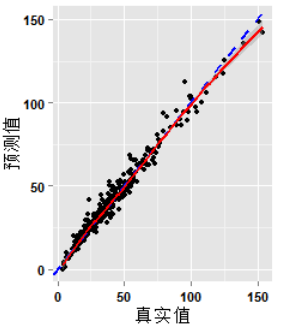
其拟合参数相关系数如表#，该模型的拟合度为。可以看出，神经生物学的数据拟合结果，除了在“重要文献数”这一参数上比微生物学有更显著的相关性，其他变量拟合系数及显著性均与微生物学高度相似。

表#

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | -8.67 | 0.78 | -11.10 |  |
|  | 发文数量 | 2.51 | 0.19 | 13.27 |  |
|  | 发文期刊数 | -0.10 | 0.023 | -4.58 |  |
|  | 重要文献数 | 0.08 | 0.018 | 4.41 |  |
|  | 总被引频次 | 0.35 | 0.013 | 26.18 |  |

从图#中可以看出，拟合模型的残差分布近似为正态分布，且真实值与预测值较好的分布在直线两侧，同样表明，模型具有较高的可信度。

通过线性回归模型分析，我们对影响h指数的因素有了更深一步的了解，通过分析我们发现，虽然两个学科的数据不通，但是线性回归结果仍然十分相似，两者之间的差异主要体现在截距上，而各变的回归系数差异很小。

图#.神经科学数据模型拟合情况

## 3.1.2 *h*指数预测分析

需要注意的是，上面的工作并没有办法对h指数进行预测分析。假如我们获得了某一时刻的h指数以及其他相关参数，若要预测若干年后的h指数，还需要用新的参数进行线性回归分析。基于对前人工作的理解以及上面的分析结果，我们采用如下思路对h指数进行预测。

首先，我们选取t时刻为2000年，用这一年的h指数及相关数据，预测预测2001年、2005年以及2010年的h指数。这里在数据选取的时候，要排除掉2000年以前没有过发文记录的研究者。再数据处理的时候，不仅要考虑我们选取的研究者中2000年及以前已经发表的文献，还要考虑后来新发表的文献。因此，在计算h指数的时候，不同的年份，总的发文量数据是有区别的（表#.）。

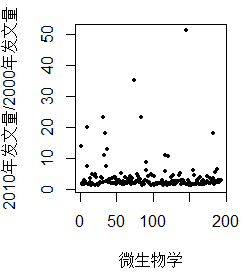
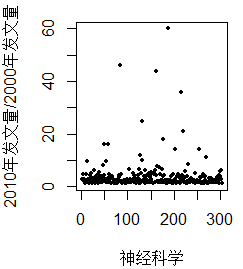
表#.不同年份的发文总数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 研究领域 | 研究者 | 2000年发文数 | 2001年发文数 | 2005年发文数 | 2010年发文数 |
| 微生物学 | 193 | 9034 | 9858 | 13388 | 18161 |
| 神经科学 | 304 | 14367 | 15526 | 20541 | 27303 |

其次，我们在分析中发现，有部分研究者刚刚从2000年以前开始发文献，而这以后的十年中，发表了大量的文献，这样我们训练集中的数据仅仅占一小部分，很明显，这样的数据将引入极端值，如图#，可以看出，大部分研究者在2000年到2010年发文量增长的倍数都相对较小，但是也存在一部分数据增长超过10倍以上。因此，我们利用2010年与2000年的发文数量之比，设定其分布的90%分位数为阈值，对数据进行进一步的筛选，筛选结果如表#.

表#.的分布情况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 研究领域 | 平均值 | 90%分位数 | 数量 | 阈值外数量 |
| 微生物学 | 3.63 | 5.23 | 173 | 20 |
| 神经科学 | 3.66 | 5.45 | 273 | 31 |

图#.微生物学以及神经科学的分布图

随后，我们分别将每一个学科的数据分为了两部分，随机抽取70%作为训练集，剩下的30%作为测试集。抽样必须保证训练集与测试集数据在分布上不存在差异，为此，我们对两部分数据进行了比较。从表#中可以看出，将数据分组之后，训练集与测试集的数据，在h指数、发文量以及总被引频次这三个参数的均值与中位数都十分接近，而且通过Kolmogorov-Smirnov检验，可以看到p值都比较大，既是说训练集与测试集的数据是同分布的，两者之间没有显著差异。

表#.训练集与测试集的数据比较

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 数据分类 | 数量 | h指数 | | 发文量 | | 总被引频次 | |
| 均值 | 中位数 | 均值 | 中位数 | 均值 | 中位数 |
| 微生物学 | 训练集 | 121 | 19.13 | 16 | 52.55 | 38 | 1828 | 904 |
| 测试集 | 52 | 18.50 | 16.5 | 49.96 | 35.5 | 1713 | 903 |
| KS检验p值 | | 0.95 | | 0.97 | | 0.89 | |
| 神经科学 | 训练集 | 191 | 20.81 | 16 | 50.69 | 34 | 2868 | 1114 |
| 测试集 | 82 | 22.18 | 17 | 56.02 | 35 | 2917 | 1118 |
| KS检验p值 | | 0.93 | | 0.66 | | 0.95 | |

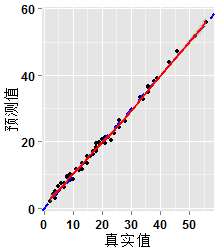
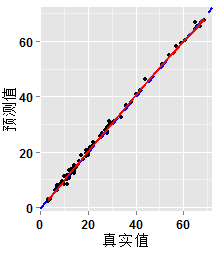
在回归分析的时候，我们选取了发文数量、发文期刊数量、重要文献数量以及总被引频次进行回归分析。另外，考虑到a值的特殊性，我们将a值也作为一个参数一起进行回归分析。综合上面的分析，我们得到了如下的回归分析结果。

首先对于微生物学的数据，预测1年后的h指数，得到的回归方程为：

拟合的相关系数如表#，从表中可以看出，对于预测1年后的h指数来说，最主要的影响因素是初始h指数，而与其他变量关系不大，该模型的拟合度。随后，我们测试集的数据进行了预测，结果如图#a，可以看出对于测试集的数据，模型的预测仍然十分准确，预测值与真实值的差异很小。

表#.微生物学1年后h指数预测拟合的相关系数表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | 0.96 | 0.44 | 2.17 | 0.032 |
|  | 初始h指数 | 0.91 | 0.043 | 21.05 |  |
|  | 发文数量 | 0.15 | 0.14 | 1.05 | 0.30 |
|  | 发文期刊数 | 0.012 | 0.018 | 0.65 |  |
|  | 重要文献数 | 0.039 | 0.027 | 1.44 |  |
|  | 总被引频次 | 0.024 | 0.019 | 1.24 |  |
|  | a值 | -0.021 | 0.050 | -0.43 | 0.67 |

图#.1年后h指数预测值与真实值的分布

随后，我们利用同样的办法，对微生物学的数据进行了预测分析，得到的1年后的h指数回归模型如下：

各参数拟合的相关系数如表#，该方程的拟合度为.可以看出，与微生物学的拟合方程相比，并无太大差异，并且在解释变量的显著性（p值）上也表现一致。同样，我们利用该模型对测试集的数据进行了预测分析，结果如图#b，可以看出，预测值与真实值分布一致。

表#.神经科学1年后h指数预测拟合的相关系数表

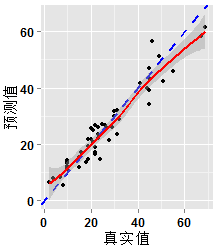
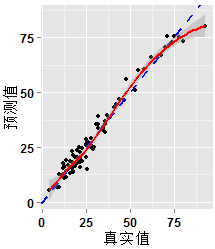
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | 0.74 | 0.35 | 2.12 | 0.036 |
|  | 初始h指数 | 0.97 | 0.033 | 29.13 |  |
|  | 发文数量 | 0.04 | 0.11 | 0.38 | 0.70 |
|  | 发文期刊数 | 0.021 | 0.014 | 1.45 |  |
|  | 重要文献数 | 0.0034 | 0.011 | 0.29 |  |
|  | 总被引频次 | 0.016 | 0.014 | 1.15 |  |
|  | a值 | 0.021 | 0.036 | 0.57 | 0.57 |

从上面的分析可以看出，对于预测1年后的h指数，其主要影响因素为初始h指数，而受其他因素影响较小。随后，我们又对5年后的h指数进行预测分析，首先在微生物学领域，我们得到了如下的拟合方程：

拟合相关系数如表#，该模型的拟合度为。与预测1年的h指数相比，可以看出，预测5年的h指数，仍然主要受初始h指数的影响，但是发文数量、重要文献数以及总被引频次也对其产生了一定影响。测试集的预测值与真实值的分布情况如图#a，与预测1年h指数相比，准确率有所下降，当h指数偏小以及偏大的时候，出现了偏离。

表#.微生物学5年后h指数预测拟合的相关系数表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | 4.67 | 1.40 | 3.34 | 0.0011 |
|  | 初始h指数 | 0.58 | 0.14 | 4.23 |  |
|  | 发文数量 | 1.00 | 0.45 | 2.23 | 0.028 |
|  | 发文期刊数 | -0.0071 | 0.057 | -0.12 |  |
|  | 重要文献数 | 0.19 | 0.086 | 2.20 |  |
|  | 总被引频次 | 0.12 | 0.060 | 2.04 |  |
|  | a值 | -0.26 | 0.16 | -1.65 | 0.10 |

图#.5年后h指数预测值与真实值的分布

在神经科学领域，我们得到了相似的结果：

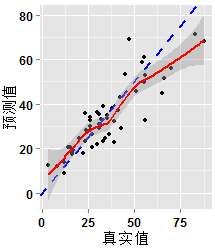
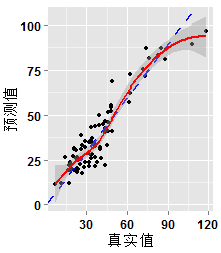
模型的拟合度为，其拟合相关系数与微生物学相似，测试集的预测值与真实值的分布情况如图#b，同样可以看出准确率稍有下降。

最后，我们又对10年后的h指数进行了预测，在微生物学领域得到了如下拟合方程：

该模型的拟合度，拟合参数的相关系数如表#，可以看出，当预测10年后的h指数是，初始h指数起的作用很弱，此时的h指数受发文数量、重要文献数的影响比较大。测试集的预测值与真实值分布如图#a，可以看出，预测的准确度不是很好。

表#.微生物学10年后h指数预测拟合的相关系数表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | 9.68 | 2.71 | 3.57 | 0.00052 |
|  | 初始h指数 | 0.30 | 0.27 | 1.11 |  |
|  | 发文数量 | 2.12 | 0.87 | 2.45 | 0.016 |
|  | 发文期刊数 | -0.028 | 0.11 | -0.25 |  |
|  | 重要文献数 | 0.37 | 0.17 | 2.24 |  |
|  | 总被引频次 | 0.17 | 0.12 | 1.48 |  |
|  | a值 | -0.51 | 0.31 | -1.68 | 0.096 |

图#.10年后h指数预测值与真实值的分布。a为微生物学，b为神经科学。

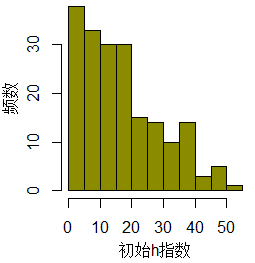
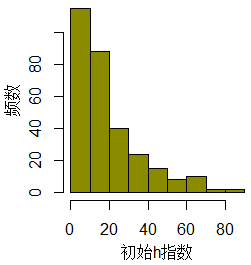
同样，我们队神经科学领域，10年后的h指数进行了预测，得到如下拟合方程：

模型的拟合度，拟合参数的相关系数如表#，可以看出，与微生物学的拟合方程相比，初始h指数对于神经科学10年后的h指数仍有较大影响。

表#.神经科学10年后h指数预测拟合的相关系数表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数 | 对应变量 | 拟合系数 | 标准差 | t值 | p值 |
|  | （截距） | 6.51 | 2.51 | 2.59 | 0.010 |
|  | 初始h指数 | 0.72 | 0.23 | 3.03 |  |
|  | 发文数量 | 1.46 | 0.79 | 1.85 | 0.066 |
|  | 发文期刊数 | 0.13 | 0.10 | 1.30 |  |
|  | 重要文献数 | 0.14 | 0.084 | 1.73 |  |
|  | 总被引频次 | 0.038 | 0.099 | 0.39 |  |
|  | a值 | 0.18 | 0.26 | 0.69 | 0.49 |

我们通过对1年后、5年后以及10年的h指数预测分析可以发现，随着时间的延长，模型的拟合度降低，同时预测的准确率也下降，这其中主要是因为数据中不同的研究者存在差异。虽然我们对发文增长的倍数进行了限制，但是我们通过分析h指数发现（图#），即使在发文增长相近的情况下，不同的研究者的初始h指数存在很大差异。对于初始h指数较小的研究者来说，在发文速度增长相近的情况下，他们的h指数增长比初始h指数较大的研究者容易。

图#.初始h指数分布。a为微生物学，b为神经科学。

综合来看，利用上面的线性拟合的方法，对一段时间之后的h指数进行预测，我们可以得到不错的预测结果，模型的拟合度较高，而且，微生物学与神经科学两个学科并未表现出太大差异。不过也必须注意到方法的局限性，对于那些开始发文量比较小，随后发文增长很快的研究者，该模型则不适用。

## 3.1.2 *g*指数与*r*指数的预测分析

从前文的分析中我们观察到，h指数、g指数以及r指数具有很强的线性关系，因此推断在预测分析的参数中，同样具有相似性，所以，我们采取了和h指数预测相同的变量，分别对g指数和r指数进行了简单的预测分析。

在微生物学领域，得到的g指数的拟合方程如下：

拟合参数系数的显著性如表#.

表#. 微生物学g指数预测模型拟合参数系数的显著性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 参数 | 对应变量 |  |  |  |
| 拟合度 | -- | 0.99 | 0.94 | 0.78 |
|  | （截距） |  |  |  |
|  | 初始g指数 |  |  |  |
|  | 发文数量 | 0.44 | 0.89 | 0.84 |
|  | 发文期刊数 | 0.027 | 0.16 | 0.13 |
|  | 重要文献数 | 0.99 | 0.43 | 0.31 |
|  | 总被引频次 | 0.44 | 0.22 | 0.16 |
|  | a值 | 0.91 | 0.99 | 0.60 |

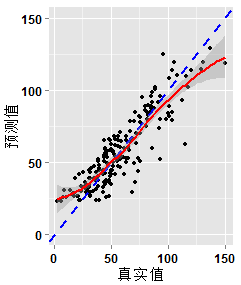
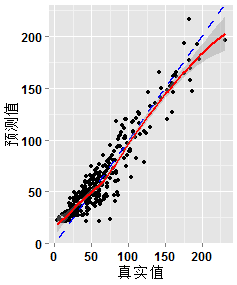
随后，我们又对神经科学领域的g指数进行预测分析，拟合方程如下：

拟合参数的显著性如表#.

表#. 神经科学g指数预测模型拟合参数系数的显著性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 参数 | 对应变量 |  |  |  |
| 拟合度 | -- | 0.99 | 0.97 | 0.88 |
|  | （截距） |  |  |  |
|  | 初始g指数 |  |  |  |
|  | 发文数量 | 0.0071 | 0.21 | 0.58 |
|  | 发文期刊数 |  |  |  |
|  | 重要文献数 | 0.93 | 0.11 | 0.051 |
|  | 总被引频次 | 0.014 | 0.029 | 0.14 |
|  | a值 | 0.61 | 0.99 | 0.64 |

从上面的结果可以看出，与h指数预测不同的是，在g指数预测的过程中，初始g指数始终对预测值有显著影响，而在神经科学领域中，发文的期刊数量对g指数产生显著影响。随后，我们对10年后的g指数预测值与真实值进行了分析，结果如图#.，从图中可以看出，与h指数预测相似，当预测10年的g指数时，准确率有所下降，而且明显可以看出，当真实g指数比较小时，预测值往往比真实值要大。

图#.10年后g指数预测值与真实值的分布。a为微生物学，b为神经科学。

随后，我们又对r指数进行了回归分析，在微生物学领域，得到了如下的拟合方程：

拟合参数的显著性如表#.

表#. 微生物学r指数预测模型拟合参数系数的显著性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 参数 | 对应变量 |  |  |  |
| 拟合度 | -- | 0.99 | 0.93 | 0.76 |
|  | （截距） |  |  |  |
|  | 初始r指数 |  |  |  |
|  | 发文数量 |  | 0.0010 | 0.13 |
|  | 发文期刊数 |  |  |  |
|  | 重要文献数 | 0.85 | 0.35 | 0.24 |
|  | 总被引频次 |  |  | 0.021 |
|  | a值 | 0.99 | 0.11 | 0.066 |

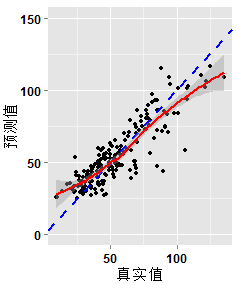
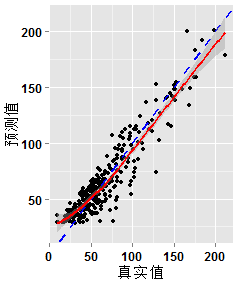
神经科学领域拟合方程如下：

拟合参数的显著性如表#.

表#. 神经科学r指数预测模型拟合参数系数的显著性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 参数 | 对应变量 |  |  |  |
| 拟合度 | -- | 0.99 | 0.96 | 0.86 |
|  | （截距） |  |  |  |
|  | 初始r指数 |  |  |  |
|  | 发文数量 |  |  | 0.0010 |
|  | 发文期刊数 |  |  |  |
|  | 重要文献数 | 0.051 | 0.050 | 0.028 |
|  | 总被引频次 |  |  |  |
|  | a值 | 0.0017 | 0.11 | 0.76 |

从拟合结果分析中可以看出，r指数与h指数、g指数最大的不同之处在于，初始r指数对于r指数预测的影响十分微弱，r指数的预测主要受到发文数量、发文期刊数以及总被引频次的影响。图#.显示了10年后r指数的预测值与真实值的分布情况，微生物学与神经科学的数据并未出现太大差异。

图#.10年后r指数预测值与真实值的分布。a为微生物学，b为神经科学。

正如我们所推断的那样，我们利用同样的方法分别对g指数和r指数进行预测分析的时候，得到的结果与h指数有很多相似性，对于h指数和g指数来说，预测中的显著性因素为初始的h指数和g指数，但是初始的r指数却对r指数预测的影响显著。在r指数预测中，总被引频次对预测结果具有显著影响。在预测中，伴随着时间的延长，模型中各个参数的显著性也发生变化。例如对预测h指数来说，随着时间延长，初始h指数对预测值不在起决定性作用；而r指数预测中，随着时间的延长，发文期刊的分布对预测结果的影响则越来越显著。

# 3.2基于引用动力学模型的探究

上文中，我们利用线性回归模型，实现了对h指数、g指数以及r指数的预测分析，但是这些指数均是从总体上描述属于同一个研究者的一群文献其以后的科学影响力的走势。在预测过程中，也没有涉及到单篇文献的被引用的趋势，因此，我们无法通过对这些指数的预测分析，了解单篇文献的引用变化情况。而利用引用动力学模型，则可以很好的解决这一问题。

引用动力学模型的表达式为：

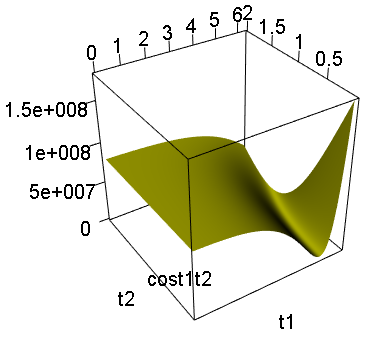
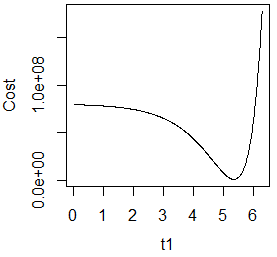
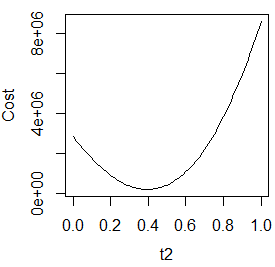
方程中，表示t时刻文献的被引频次，对于不同的文献来说，我们只需要求出模型中的,与，既可以对单篇文献进行拟合分析。这里我们主要采用最小二乘法进行拟合分析，利用梯度下降的方法求出参数的最优值，其研究思路如下。

首先我们假设函数为要拟合的函数，那么模型拟合的成本函数(Cost Function)为：

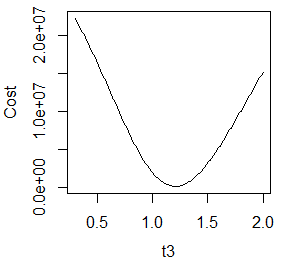
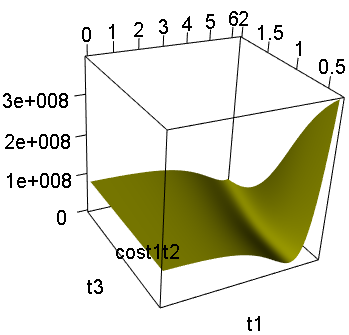
我们只需要通过改变的值，从而使最小化。为了实现这一目标，我们首先需要确定，成本函数随的变化是收敛的。而对于引用动力学模型来说，对应的分别为以及，我们随机选取两篇文献，来验证成本函数关于以及是否具有收敛性，其中具体被引频次如图。我们对文献的累计被引频次变化过程分析发现，所有的文献几乎都遵循图中的增长规律，刚开始获得的被引的速率较快，随后增长逐步变慢。当然也有例外，在引文分析中有一种现象叫做“睡美人”文献，指的是文献的被引频次在沉寂了一段时间之后，突然出现爆发性的增长，这种特殊现象不在我们的研究范围内。

由于存在以及三个参数，为了能够直观的看出成本函数随参数的变化情况，我们采用固定一个参数，然后两两分析参数变化与成本函数的关系，从而验证成本函数是否具有收敛性。首先，我们根据Barabási的研究结果，固定，分析与对成本函数的影响，经过调整，得到结果如图#。从图中可以看出成本函数随着t1与t2的变化，在一定范围内，会出现最小值。随后我们又分析了成本函数随t1与t3变化的情况得到的结果十分相似，同样，成本函数随t2与t3变化的情况也具有收敛的特征。

图#.累计频次随年份的变化

图#.成本函数随t1和t2变化的关系



图#.成本函数随t1和t3变化的关系

第四章 关于科学影响力评价与预测的思考

# 4.1论文总结

# 4.2局限性与展望

1. 公式的R语言表达
2. 函数的R语言表达

参考文献

致谢

攻读学位期间发表的学术论文

1. 数据来源网站：<http://f1000.com/prime/thefaculty/microbiol> [↑](#footnote-ref-1)