基于聚类与时间序列的H指数增长预测分析

李子岳 肖昌荣

清华大学经济管理学院

2020 年 5 月 21 日

目录

H指数简介

- Hirsch 于 2005 年提出。
- 目的：量化学者的影响力和科研成果，进行评估比较。
- H-index $=$ h 表示
  - 至少有 h 篇论文满足每篇论文至少被 h 篇论文引用。
  - 而对于该学者其他的论文，每篇论文引用量则小于等于 h。
- 是目前广为学界接受的衡量学者学术影响力的重要指标。

H指数的预测

- “学术年龄”的重要性 (*Penner et al., 2013*)
  - “学术年龄”即从第一篇文章发表至今的年数。
  - H 指数是一个随着时间积累的数，与学术年龄有很强的相关性。但在预测时需要消除 H 指数随时间积累的性质。
  - → 整合移动平均自回归模型 (ARIMA 模型)
- 基于回归的预测方法 (*Dong et al., 2016*)
  - 回归因子：当前的 H 指数值、发表的文章数量、总引用量、合作者的数量、从第一篇文章发表至今的年数。
  - 在本文数据集中的表现不佳：回归因子中仅“当前的H指数值”一项的参数为显著。
  - → 基于聚类的预测模型

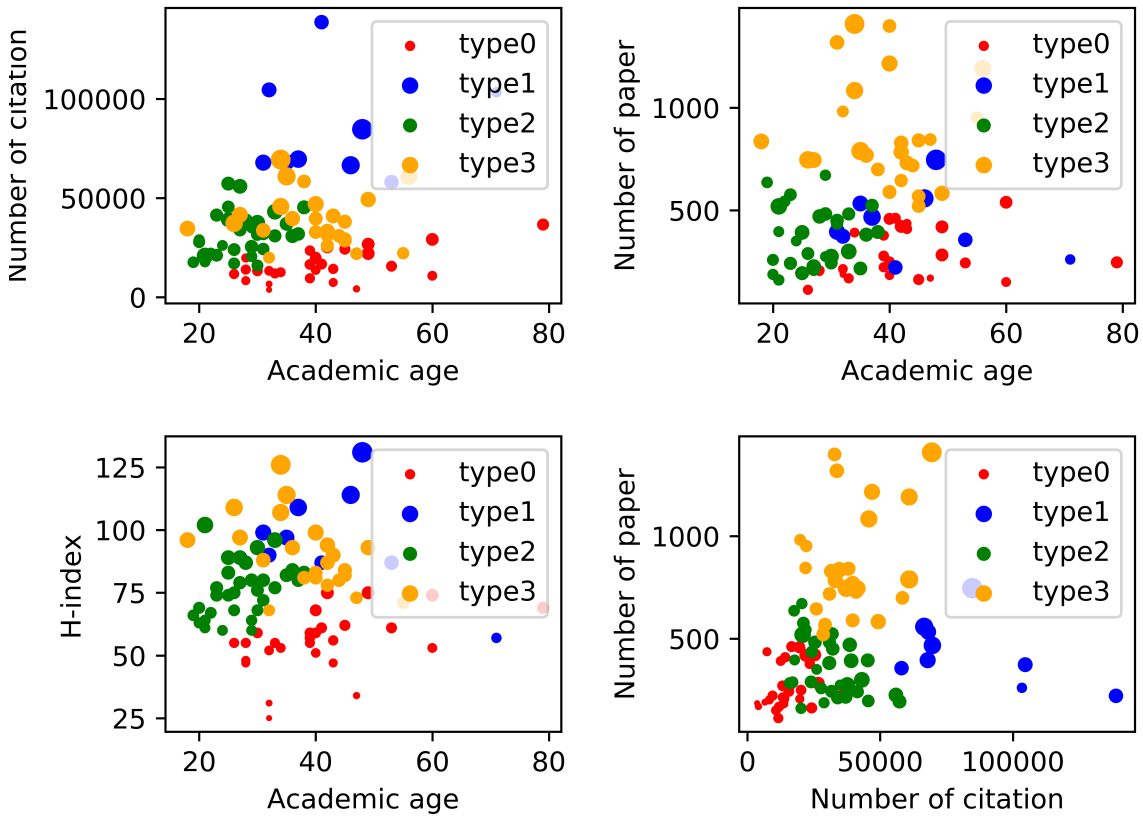
数据集简介

- 从 Scopus 中收集了 96 位计算机科学领域的知名学者的数据。

名称	解释
h-index	从1970年至2020年该学者截止每年的h指数值。
earliest_year	该学者发表最早一篇文章的年份。
academic_age	该学者的“学术年龄”。
co-author	该学者截止2020年的合作者数。
paper_num	该学者截止2020年发表的文章数。
citation_num	该学者截止2020年的总被引用量。
journal_num	该学者截止2020年的所有发表文章所在期刊数。

聚类分析

- 将数据集中的 H 指数值、学术年龄、发表文章数量、总引用量作为每位学者的特征进行聚类。
- 使用 *K-means* 方法，确定最优聚类类别数 $k = 4$。

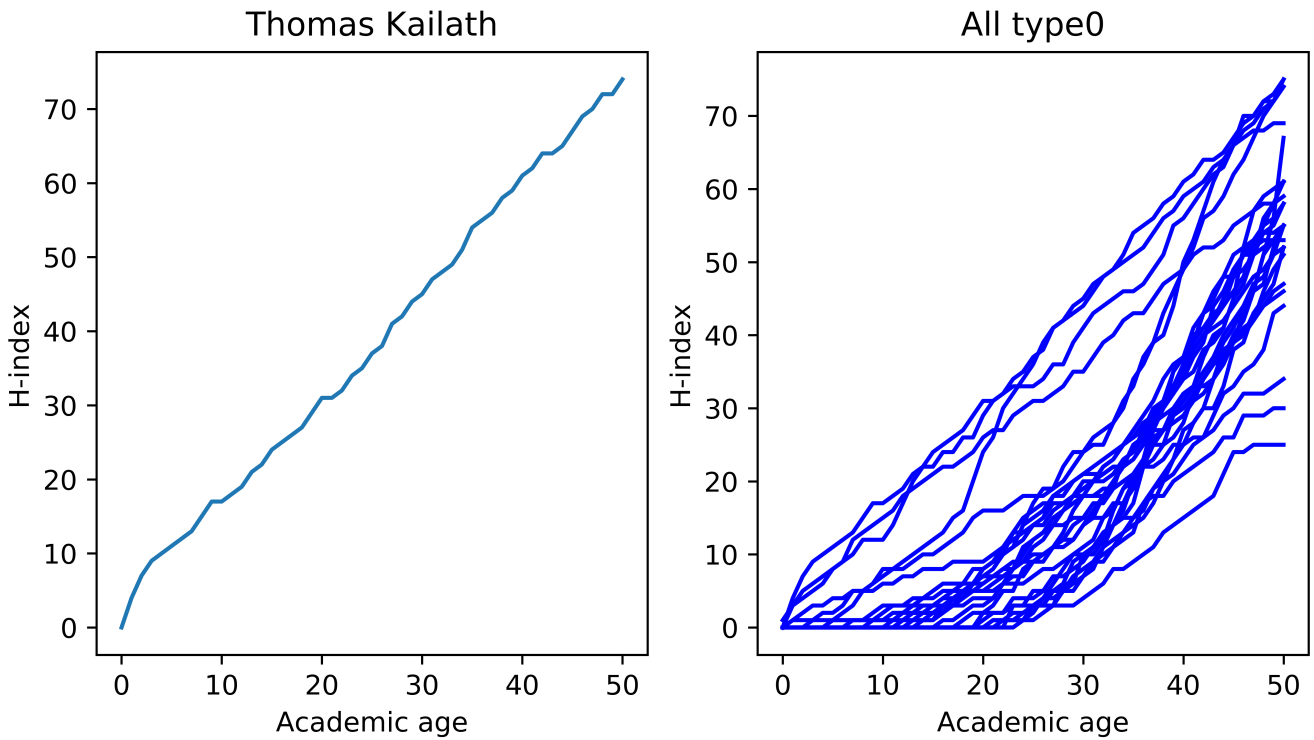


聚类分析 (Cont'd)

- Type0: 较弱势学者
  - 分布于各个学术年龄段，无论是发表文章数、总引用量，还是 H 指数都属于四个类别学者中最低的一个。
  - 可能原因：研究方向不热门、中国学者。
- Type1: 引用量优势学者
  - 虽然发表文章数不多，但总引用量是最高的，进而 H 指数值也普遍很高。
- Type2: 年轻型学者
  - 学术年龄相对而言是最年轻的，发表文章数、总引用量、H 指数值中等，有较大的发展潜力。
- Type3: 文章量优势学者
  - 总引用量不是最高，但凭借高发表文章数得到较大的 H 指数。

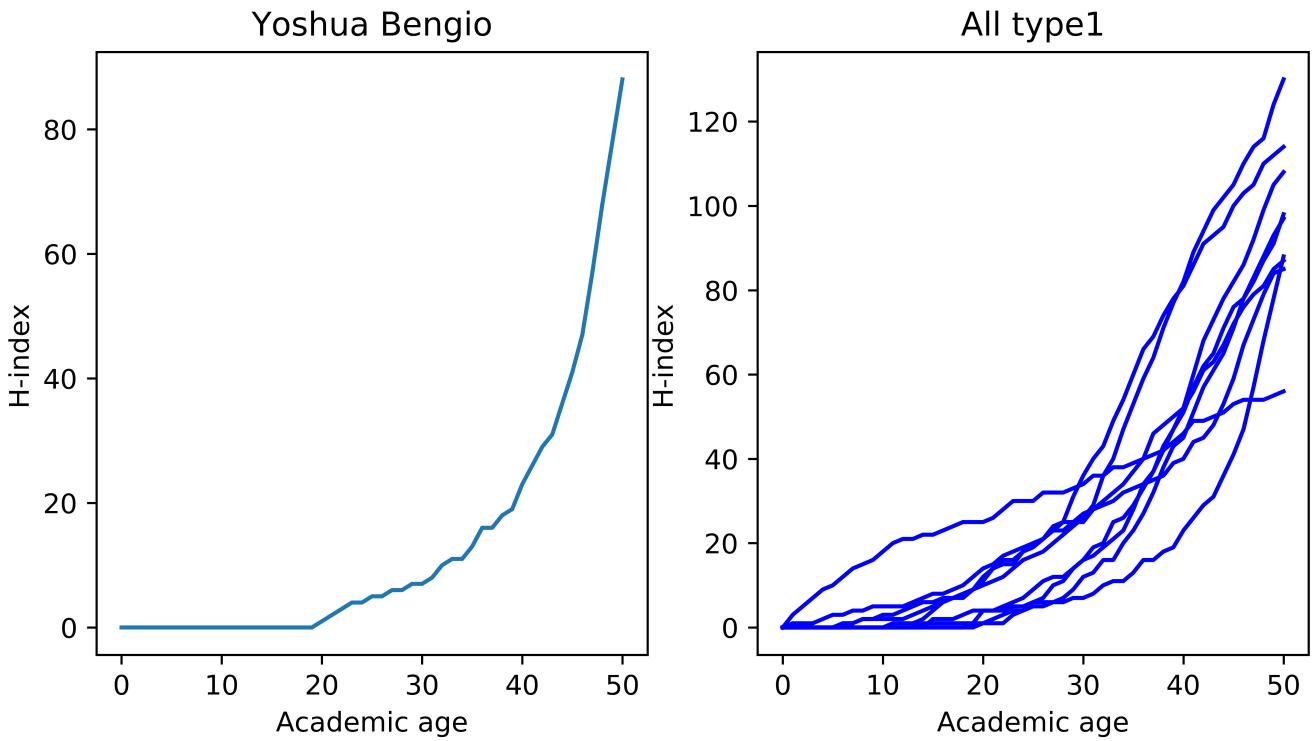
H指数增长规律分析

- 较弱势学者 $\rightarrow$ 线性模型
  - 举例：Thomas Kailath。线性系统领域，较为冷门。



H指数增长规律分析 (Cont'd)

- 年轻型学者、引用量优势学者、文章量优势学者 → 指数模型
  - 举例：Yoshua Bengio。引用量优势学者，人工神经网络和深度学习领域专家，2018年图灵奖获得者。

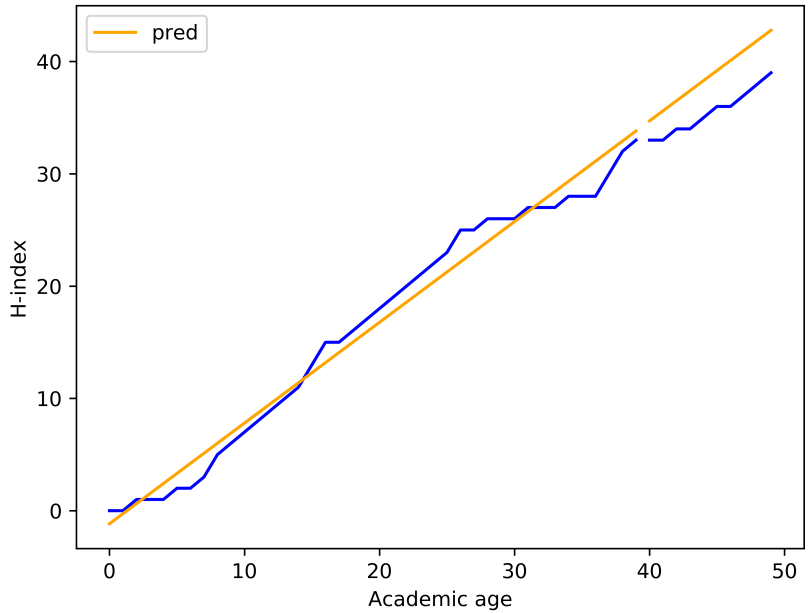


模型构建

- 思路：通过聚类结果确定学者 H 指数增长模式。
- 基本步骤：
  1. 利用学者的 H 指数历史数据、当前学术年龄、发表文章数量、总引用量等特征，对数据集中的学者进行预先聚类。
  2. 将所要预测的学者归类（如选取与其欧式距离最近的聚类中心所在类别）。
  3. 根据步骤2中的类别确定其 H 指数增长模式（较弱势学者对应线性模型，其他三类学者对应指数模型）。
  4. 拟合增长模型，得到模型参数。
  5. 利用模型进行 H 指数值的预测。

预测实例

- 选取清华大学姚期智教授的数据，依据模型构建部分所述的聚类启发式模型构建方法构建模型。归类结果为较弱势学者，应用线性增长模型。
- $\hat{\beta}_1 = 0.8969$，在 99% 的置信水平显著。$R^2 = 0.981$。



模型局限性

- 聚类启发式模型的必要性与有效性。
- 依赖于多种数据，数据质量关系到模型预测的准确性。
- 只能进行短期预测。因为聚类的特征中包含了学术年龄，一些类别的模型仅对一定年龄段的学者有效。
- 只能应用在同一或相似的学科领域。
- 对学术年龄非常年轻的学者进行预测的效果不好。

总之，聚类启发式模型的思想是充分利用学者自身的多种数据、其他学者的 H 指数增长模式等多维数据提供的信息精细化地确定增长模式。目前的模型仅为可供参考的想法，有很大的改善空间。

模型构建

- 确定 ARIMA(p, d, q) 的阶数
- 使用 AIC 准则

表: 不同p, d, q出现频次

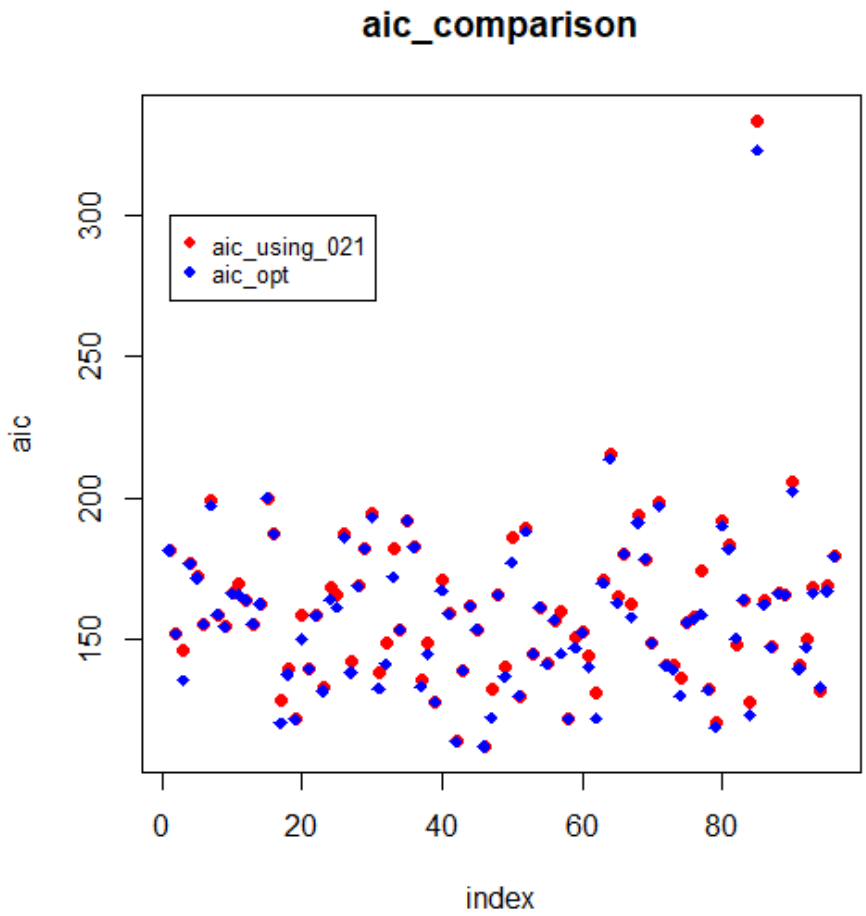
p	d	q	frequency
0	2	0	7
0	2	1	42
0	2	2	6
1	2	0	15
1	2	1	2
1	2	2	3
2	2	0	4
2	2	1	1
2	2	2	9
other parameter orders			7
total			96

ARIMA(0, 2, 1) 模型：

$$\nabla^2 h_t = (1 + \theta B)\epsilon_t$$

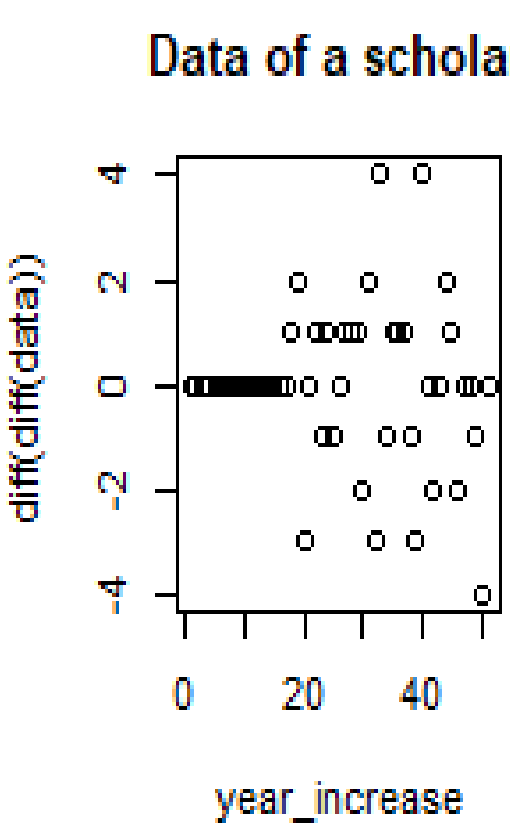
模型构建(Cont'd)

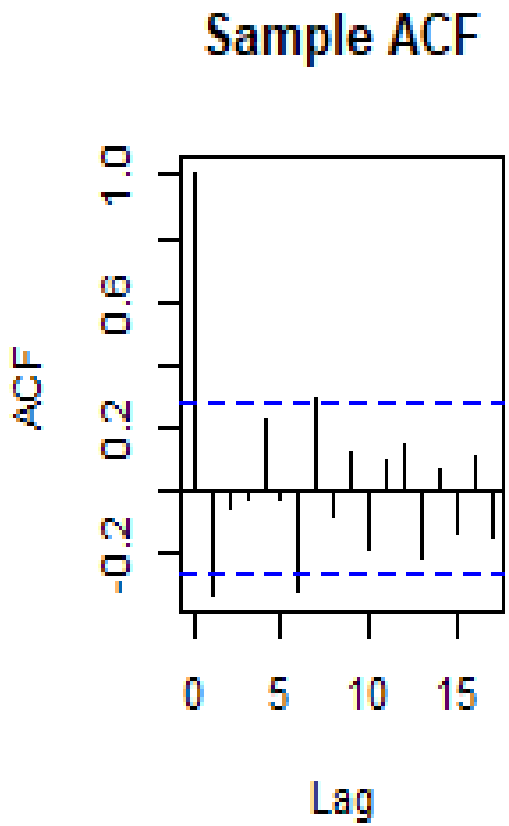
ARIMA(0,2,1) 的 AIC 与使 AIC 最小的 ARIMA 模型的 AIC_{opt} 的比较

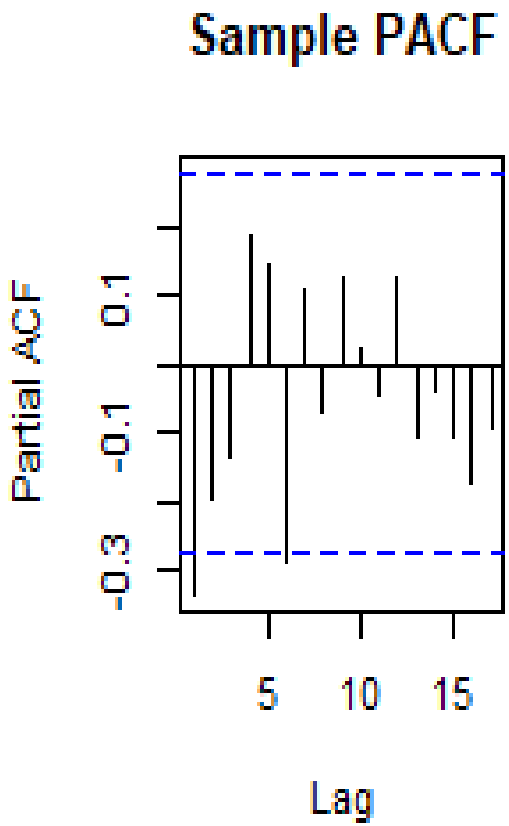


模型构建(Cont'd)

- 实例分析
- 平稳性、ARIMA模型定阶

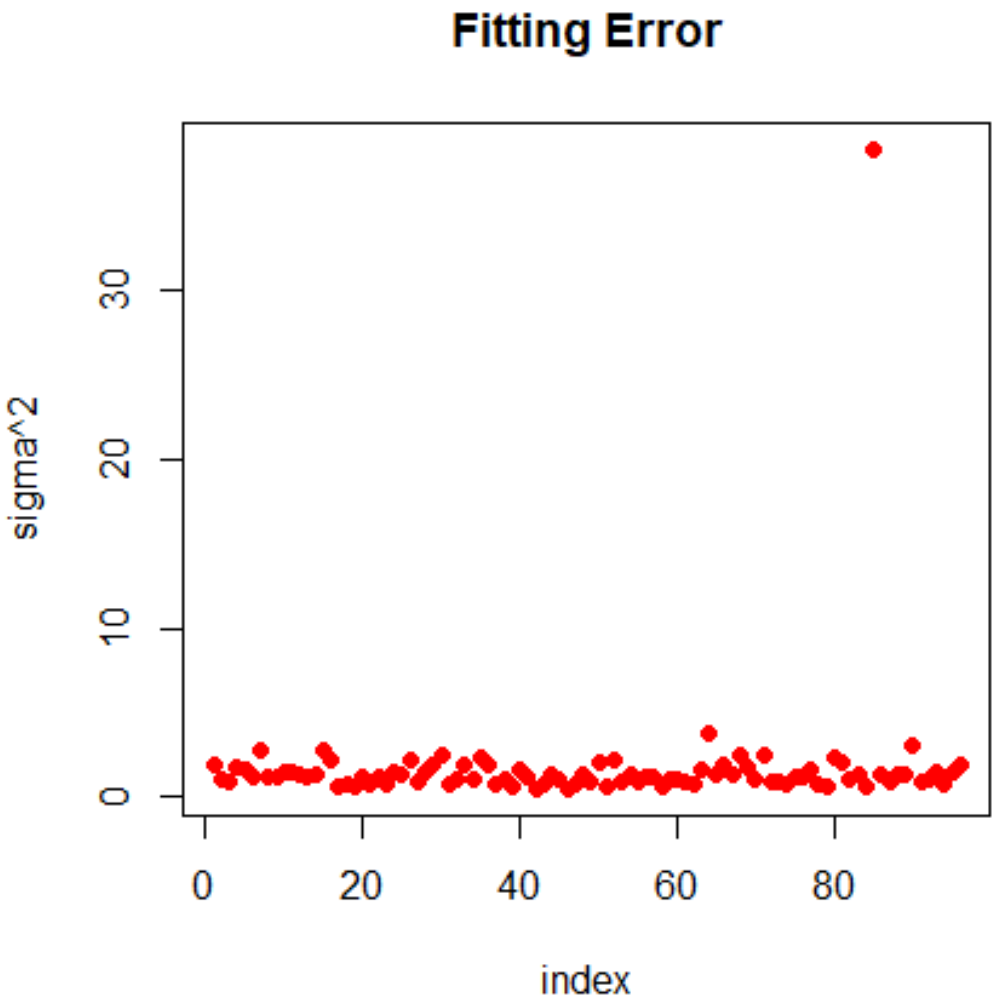






模型拟合

去除离群值之后，σ^2的估计值的平均值为1.358，方差为0.359。



预测效果

- 使用每一个学者前 i 年的数据作为已知数据，由此估计模型中的未知参数，并对第 $i+n$ 年的 H 指数进行预测。
- 随着 n 的增加，预测误差增加。
- 模型有低估未来 H 指数的倾向。

表: 预测误差的均值

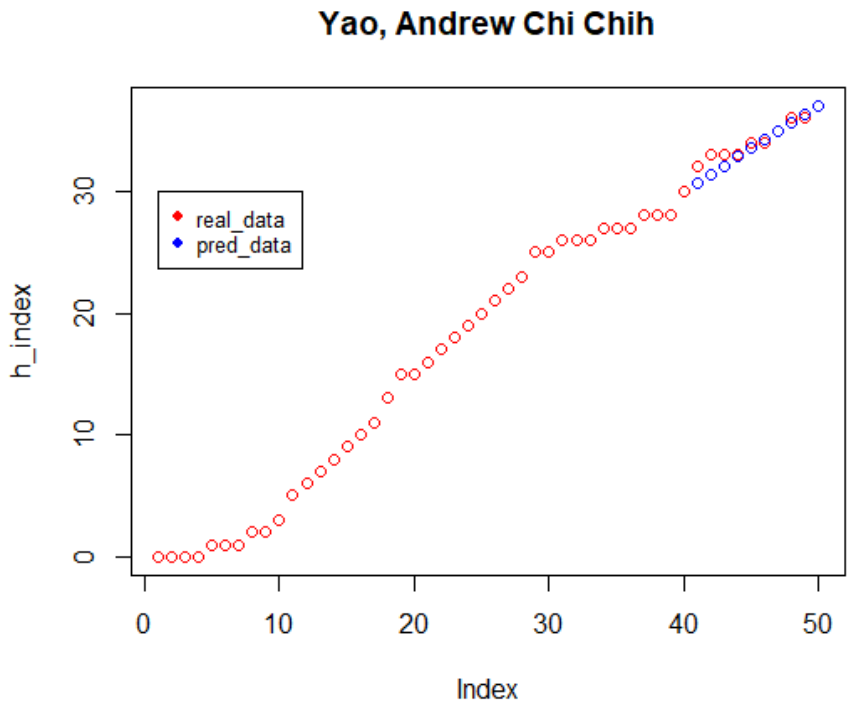
$n\backslash i$	36	37	38	39	40	41	42	43	44	45
1	0.5637	0.6073	0.0047	0.4508	0.126	0.4356	0.0412	0.0581	0.1726	0.4144
2	1.4398	0.9125	0.4364	0.7869	0.5957	0.6733	0.1346	0.2724	0.6368	0.9226
3	2.0139	1.6448	0.7536	1.4668	0.8675	0.9631	0.3841	0.7784	1.1947	1.1911
4	3.0151	2.2625	1.4145	1.9488	1.1913	1.4091	0.9254	1.3782	1.5131	1.3451
5	3.9016	3.2239	1.8776	2.4829	1.6715	2.1468	1.5603	1.7383	1.7169	1.3324

表: 预测误差的方差

$n\backslash i$	36	37	38	39	40	41	42	43	44	45
1	2.2576	1.8125	1.9591	3.1763	1.6137	1.9155	1.4862	2.4509	2.0032	3.2183
2	6.1407	3.7774	6.7228	8.9908	4.737	5.0371	4.7586	6.3608	7.1244	25.021
3	10.179	8.6098	15.388	17.016	8.7592	10.763	9.5659	13.091	35.449	67.909
4	18.445	16.298	27.092	27.864	15.949	18.282	15.574	43.834	85.479	90.932
5	30.142	25.74	43.691	46.913	24.859	26.258	48.36	95.252	112.86	113.09

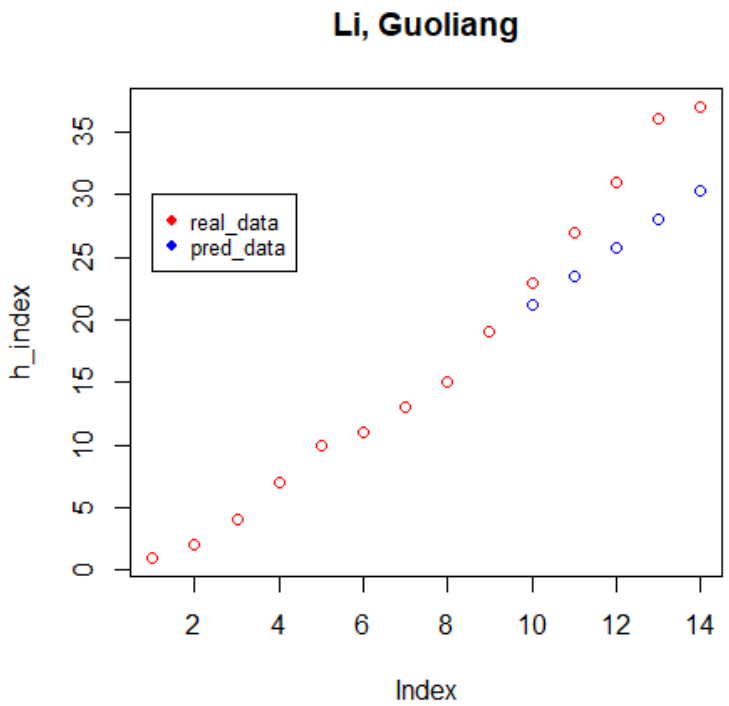
预测实例

- 对清华大学姚期智教授 H 指数的预测。
- 前40年的数据为已知，对后10年的数据进行预测。
- 预测值（蓝色数据点）与真实值十分接近!



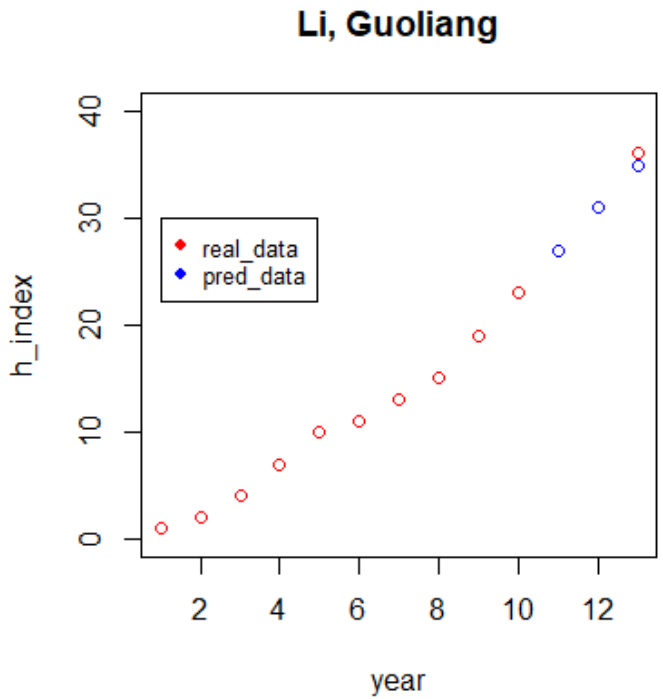
预测实例(Cont'd)

- 对清华大学李国良教授 H 指数的预测。
- 李国良教授比较年轻，成长速度过快，模型低估 H 指数的增长。



预测实例(Cont'd)

- 对清华大学李国良教授 H 指数的预测。
- 若增加已知 H 指数值的数量，并减少预测的年份，预测结果更精确。



模型评估

- 对于资深学者和领域专家，预测较为准确。
- 对于年轻学者或成长速度较快的学者，预测较为保守，H指数增长偏慢——干预分析（Intervention Analysis）。
- 随着预测年份 n 的增加，预测误差增加。

模型总结

- 聚类启发式模型
  - 在学者 H 指数历史数据未知的情况下，利用同领域相似学者的增长模式进行预测。
  - 由于数据量较少，聚类结果有待提高。
- ARIMA(p,d,q)模型
  - 使用已知的 H 指数信息直接推断未来的 H 指数信息。
  - ARIMA(0,2,1) 模型对大部分学者的（较短期）预测非常精确。
  - 当学者在某一时刻H指数增长突然加快时预测偏差较大——干预分析。

H指数的局限性

- 在非主流领域工作的学者，H 指数可能相对偏低。
- 不同领域之间的 H 指数也可能存在较大的差异。
- 有的文章作者很多，一些学者的文章和引用数量因此大幅增加。
- H指数与总引用数存在较强的函数关系，Alexander(2014)给出 $h \approx 0.54\sqrt{N_{citations}}$。
- 学者“科研年龄”的决定不应该是简单的发表第一篇文章。
- 自引可以显著增加 H 指数。

H指数的改进

- 考虑第一作者、第二作者。
- 删除自引数。
- Egghe(2006) 提出了 G 指数的理论与实践。
  - 一名学者的“G 指数$=g$”表示，在他的所有论文中，引文数量最高的 g 篇论文的总引用数大于等于 g^2，满足此条件的最大的 g 即为 G 指数。
  - 解决了部分学者通过增加发表文章数量和一定数量的文章引用（包括自引）等提高 H 指数的问题，同时对于非主流领域工作的学者成就的评估也相对更为客观。

参考文献

ALEXANDER Y, 2014. Critique of hirsch's citation index: a combinatorial fermi problem[J/OL]. arXiv.org. https://search.proquest.com/docview/2084724191?accountid=14426.

BAR-ILAN J, 2008. Which h-index? — a comparison of wos, scopus and google scholar[J/OL]. Scientometrics, 74(2): 257-271. https://doi.org/10.1007/s11192-008-0216-y.

Dong Y, Johnson R A, Chawla N V, 2016. Can scientific impact be predicted?[J/OL]. IEEE Transactions on Big Data, 2(1): 18-30. DOI: 10.1109/TBDATA.2016.2521657.

EGGHE L, 2006. Theory and practise of the g-index[J/OL]. Scientometrics, 69(1):131-152. https://doi.org/10.1007/s11192-006-0144-7.

HIRSCH J E, 2005. An index to quantify an individual's scientific research output[J/OL]. Proceedings of the National Academy of Sciences, 102(46):16569-16572. https://www.pnas.org/content/102/46/16569. DOI: 10.1073/pnas.0507655102.

PENNER O, PAN R K, PETERSEN A M, et al., 2013. On the predictability of future impact in science[J/OL]. Scientific Reports, 3(1):3052. DOI: 10.1038/srep03052.

Q&A