定量研究方法（三）

文本计量法

林景 讲师

南京财经大学法学院

为什么需要文本计量法

- 文本计量法：把蕴含各种信息的书面材料和文字材料作为研究对象的定量研究方法。
- 文本（literature）的类型：

	初级/一手的（first-hand）	次级/二手的（second-hand）
现时的	研究者亲自记录的具有时效性的文本，如法庭记录、会议记录、调查记录、录音记录等	研究者对别人的记录进行汇编的具有时效性的文本，如庭审总结、年度总结、调查总结、采访总结等
回顾的	研究者亲自记录的不具有时效性的文本，如回忆录、传记、观后感、日记、回溯报告等	研究者对别人的记录进行汇编的不具有时效性的文本，如地方志、读书笔记、墓志铭、历史总结等

为什么需要文本计量法

- **文本作为研究对象**的定量研究：
  - 研究对象无反应性
  - 费用低、省钱省时
  - 可以研究那些无法接触的研究对象
  - 适合做纵贯分析
  - 研究保险系数较大

为什么需要文本计量法

- **文本作为研究对象**的定量研究：
  - 相比历史学研究的优势
    - 将史料与社会科学理论相结合
    - 将史料与定量分析技术相结合
  - 相比问卷调查法和实验法的优势
    - 将海量历史社会现象重新带回社会科学研究
    - 可以进行跨学科、跨区域和跨时期比较研究

文本计量法案例：中古中国的门阀大族为何消亡？

- 郑樵（1104-1162），宋代史学家
  - “自隋唐而上，官有簿状，家有谱系。官之选举，必由于簿状；家之婚姻，必由于谱系……自五季以来，取士不问家世，婚姻不问阀阅”
  - 关陇贵族：北朝延续自隋唐时期的关中、陇西门阀贵族
  - 唐朝魏博节度使：田承嗣→田悦→田绪→田季安→田怀谏

文本计量法案例：中古中国的门阀大族为何消亡?

- 历史学的解释：唐宋变革论
  - 唐代是中国中古时代的结束，宋代是中国近世时代的开始；
  - 唐代的大部分政治精英（political elites）可追溯自数百年以前的未曾断裂的家族系谱，但这些政治精英家族在宋体突然消亡；
  - 唐宋变革是中国社会政治精英的性质和构成的重大转变：
    - 贵族阶层在政治动荡和权力斗争中失势，政治权力逐渐集中在君王中心；
    - 科举制加速的社会阶层的流动，平民有机会进入官僚士大夫阶层；
    - 土地政策的变革使土地不再垄断在贵族手中，而是分散到平民阶层。

文本计量法案例：中古中国的门阀大族为何消亡?

- 谭凯（Nicolas Tackett），加州大学伯克利分校
  - 数据来源：墓志铭
    - 作为随葬品的墓主传记
    - 晚唐诗文别集、《旧唐书》、《新唐书》的人物传记
    - 地方碑林、神道的墓志拓片
  - 数据编码
    - 墓主的出生与死亡时间
    - 墓主的身份、阶层、地位及其变动
    - 墓主的配偶、子女、父母以及其他家族成员身份

文本计量法案例：中古中国的门阀大族为何消亡?

- 谭凯（Nicolas Tackett），加州大学伯克利分校
- 黄巢之乱彻底冲垮了中国的贵族体系
  - 过去政治动荡和权力斗争没有对中国贵族体系造成根本改变，门阀大族之间通过政治联姻重新适应新的社会政治体制，继而复兴；
  - 科举制对平民阶层向上流动作用有限，唐代通过科举做官的群体主要来自门阀大族。他们在科举竞争中占据绝大部分优势；
  - 公元880年，以盐商黄巢为首的地方民变势力崛起，血腥清洗了主要定居在长安、洛阳以及两都之间的两京走廊地带的门阀大族，才造成后者衰亡。

国外文本计量数据库

- 中国历代人物传记资料库（China Biographical Database Project）
  - 哈佛大学、北京大学、台北“中央研究院”
  - https://projects.iq.harvard.edu/cbdb
- 中国近3000气候纲要数据库（Compendium of Meteorological Records of China in the Last 3000 Years ）
  - 威斯康辛大学麦迪逊分校
  - https://www.nature.com/articles/sdata2018288
- 中国精英与金融政治数据库
  - 加州大学圣地亚哥分校
  - https://gps.ucsd.edu/news-events/news/confronting-china-elite-and-financial-politics.html

国内文本计量数据库

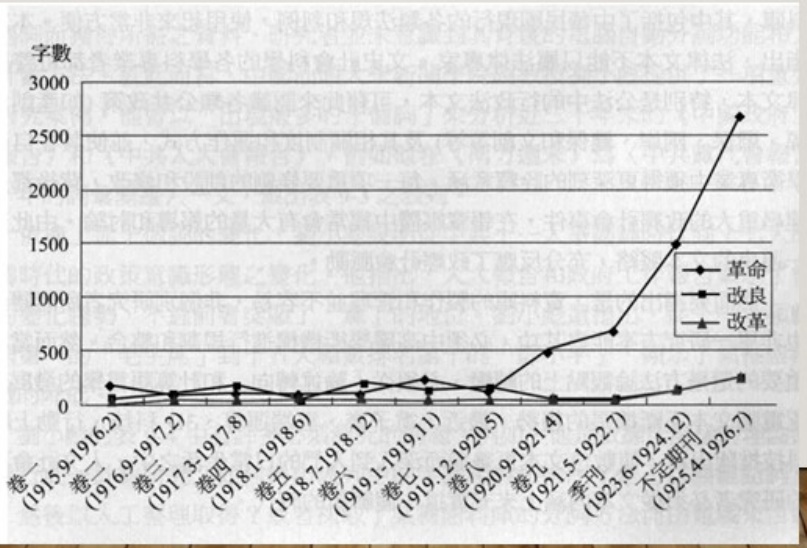
- 中国历史官员量化数据库（清代）
  - 香港科技大学社会科学院
  - http://vis.cse.ust.hk/searchjsl/
- 中共政治精英资料库
  - 台湾政治大学政治学系
  - https://cped.nccu.edu.tw/
- 中国方言数据库和中国资本主义工商业改造历史数据
  - 中山大学岭南学院产业与区域经济研究中心
  - https://gitee.com/arlionn/IRE
- 中国地方历史文献数据库
  - 上海交通大学
  - http://dfwx.datahistory.cn/pc/

国内文本计量数据库

- 中国近现代思想及文学史专业数据库（1830-1930）
  - 台湾政治大学
  - http://dsmctl.nccu.edu.tw/d_about.html
- 中国近代思想史专业数据库
  - 香港中文大学中国研究中心
  - https://www.cuhk.edu.hk/ics/rcccc/database_main.html

文本计量法可以做什么？

- 描述分析：文本变量的描述统计
- 金观涛，《观念史研究：中国现代重要政治术语的形成》
  - 《新青年》杂志中的政治术语流变



文本计量法可以做什么？

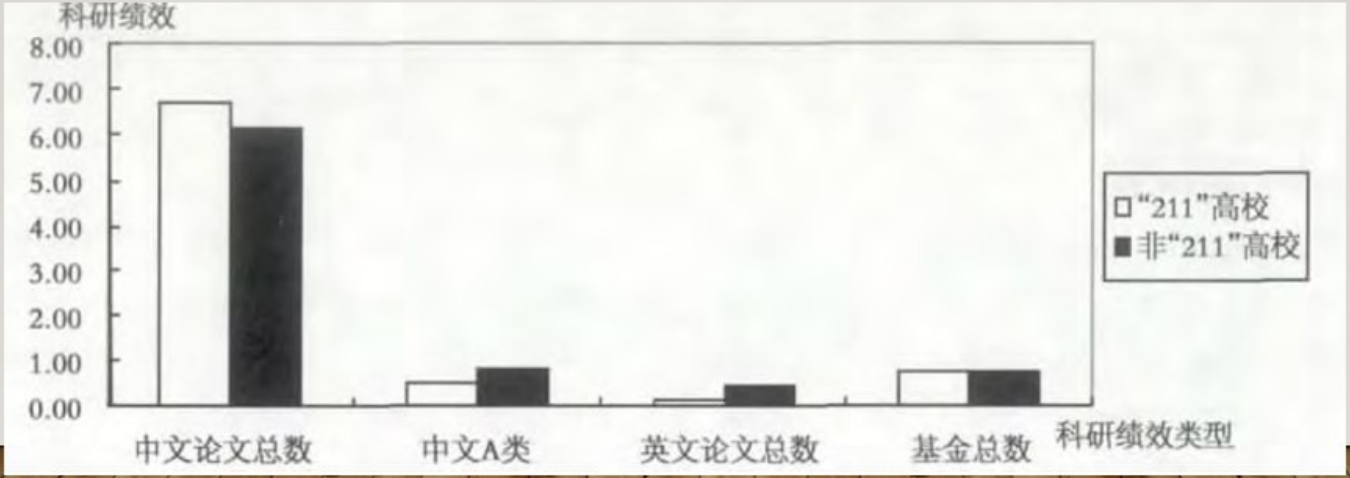
- 描述分析：文本变量的描述统计
- 王军辉等，《科研人才招聘需要“查三代”吗？》
  - 16所经济学院中的教育背景与科研绩效

表1　描述性统计量

变　量	观测值	均值	标准差	最小值	最大值
个人特征					
性别	1 036	0.292	0.455	0	1
是否留校	1 036	0.546	0.498	0	1
最后学位信息					
博士(或最后学位)毕业时间	1 036	2001	7	1981	2012
博士(或最后)毕业学校排名	1 036	5.084	1.967	1	9
是否博士	1 036	0.874	0.333	0	1
是否留学	1 036	0.255	0.436	0	1
是否出国访问	1 036	0.281	0.450	0	1
本科学习信息					
本科“211”高校	1 036	0.732	0.443	0	1
本科“985”高校	1 036	0.629	0.483	0	1
本科“211”但非“985”	1 036	0.102	0.303	0	1
经管类专业	1 036	0.580	0.494	0	1
理工类专业	1 036	0.278	0.448	0	1
其他专业(文法医等)	1 036	0.105	0.307	0	1

文本计量法可以做什么？

- 描述分析：文本变量的描述统计
- 王军辉等，《科研人才招聘需要“查三代”吗？》
  - 16所经济学院中的教育背景与科研绩效



文本计量法可以做什么？

- 因果关系分析：文本的变量间关系
- 王军辉等，《科研人才招聘需要“查三代”吗？》
  - 16所经济学院中的教育背景与科研绩效

最后学位毕业时间	最后毕业时间 > 1980			最后毕业时间 > 1990			最后毕业时间 > 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
变量	年均中文论文数	年均中文A类论文数	年均英文论文数	年均中文论文数	年均中文A类论文数	年均英文论文数	年均中文论文数	年均中文A类论文数	年均英文论文数
				全样本					
*BA*211	0.039	0.009	0.040*	0.051	0.009	0.045*	−0.019	0.015	0.060
	(0.066)	(0.014)	(0.021)	(0.075)	(0.016)	(0.022)	(0.061)	(0.020)	(0.035)
*BA*985	−0.081	0.010	0.026	−0.069	0.010	0.027*	−0.065	0.020	0.032*
	(0.063)	(0.010)	(0.015)	(0.071)	(0.010)	(0.014)	(0.063)	(0.012)	(0.017)
性别	−0.240***	−0.044***	−0.013	−0.225***	−0.044***	−0.014	−0.177***	−0.043**	−0.019
	(0.023)	(0.012)	(0.008)	(0.023)	(0.012)	(0.008)	(0.026)	(0.015)	(0.011)
留校	−0.165**	−0.055***	−0.023**	−0.140**	−0.050***	−0.027*	−0.197***	−0.057***	−0.032*
	(0.057)	(0.013)	(0.011)	(0.053)	(0.010)	(0.013)	(0.036)	(0.014)	(0.018)

文本计量法可以做什么？

- 因果关系分析：文本与其他数据库的变量间关系
- 周飞舟，《“三年自然灾害”时期我国省级政府对灾荒的反应和救助研究》

表 7　死亡率最高的 11 省四个年度的调拨粮和死亡率

	1958		1959		1960		1961	
	调拨粮	死亡率(‰)	调拨粮	死亡率(‰)	调拨粮	死亡率(‰)	调拨粮	死亡率(‰)
安徽	－721	12.36	－462	16.72	－260	68.58	37	8.11
山东	35	12.77	－33	18.14	372	23.51	312	18.49
河南	－739	12.69	－402	14.10	46	39.56	285	10.20
湖北	－222	9.60	200	14.49	－324	21.22	108	9.08
湖南	－806	11.58	－1133	12.92	－384	29.26	－220	17.48
广西	157	11.98	307	17.33	176	29.20	87	20.37
四川	－1959	17.37	－2381	19.22	－1390	47.78	－248	28.01
贵州	－241	15.26	－348	20.28	－209	52.33	30	23.27
云南	－23	21.62	－3	17.96	－52	26.26	－128	11.84
甘肃	－151	21.22	－100	17.38	67	41.32	164	11.48
青海	66	12.64	66	16.29	66	40.73	66	11.68

文本计量法的实施步骤

- 从理论出发确定经验问题和进行概念操作化
- 将文本资料转化定量数据库
  - 选择文本来源，汇总文本资料
  - 确定文本的分类范畴和分析单位
  - 确定编码体系，将文本转化为具体变量
  - 构建定量数据库
- 选择适切统计技术进行分析
- 进行信度和效度检验

文本计量法的实施步骤：以姓氏排名和学术成就为例

- Liran Einav & Leeat Yariv
  - "What's in a Surname? The Effects of Surname Initials on Academic Success"
  - *Journal of Economic Perspective*
- 理论背景：学术合作与学术贡献的悖论
  - 学术合作在经济学、社会学和政治学等社会科学领域已经成为普遍现象；
  - 在论文发表中，按**作者姓氏排名**和按**作者贡献排名**是两种主要形式，但往往**姓氏首字母在前**的学者在学术事业上更为成功；
  - 姓氏排名靠前意味着拥有更高的学术能见度（visibility）；
  - "李-杨之争"（《谁得到了爱因斯坦的办公室》，2016，上海科技教育出版社）

文本计量法的实施步骤：以姓氏排名和学术成就为例

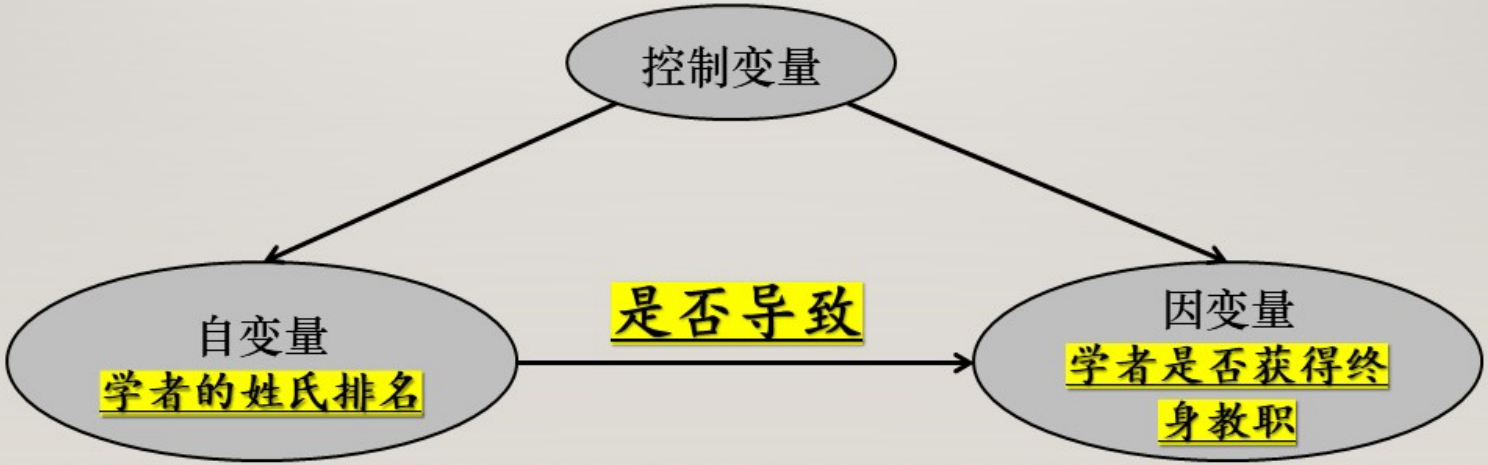
- 经验现象与研究问题
  - 资深教授相比资浅教授，姓氏排名更靠前；
  - 名牌大学教授相比一般大学教授，姓氏排名更靠前；
  - 拥有终身教职的教授比非终身教职的教授，姓氏排名更靠前；
  - **姓氏排名是否影响学者的学术成就？**

文本计量法的实施步骤：以姓氏排名和学术成就为例

- 概念操作化
- 因变量（dependent variable）/结果变量（outcome variable）
  - 学术成就：学者是否获得终身教职；
    - 定类变量，取值范围为0-1；
    - 获得终身教职编码为1，没有获得终身教职编码为0。
- 自变量（independent variable）/原因变量（causal variable）
  - 姓氏排名：学者的姓氏首字母在字母表中的位置；
    - 定距变量，取值范围为1-26；
    - 如**Abott编码为1**，**Zuckerberg编码为26**。

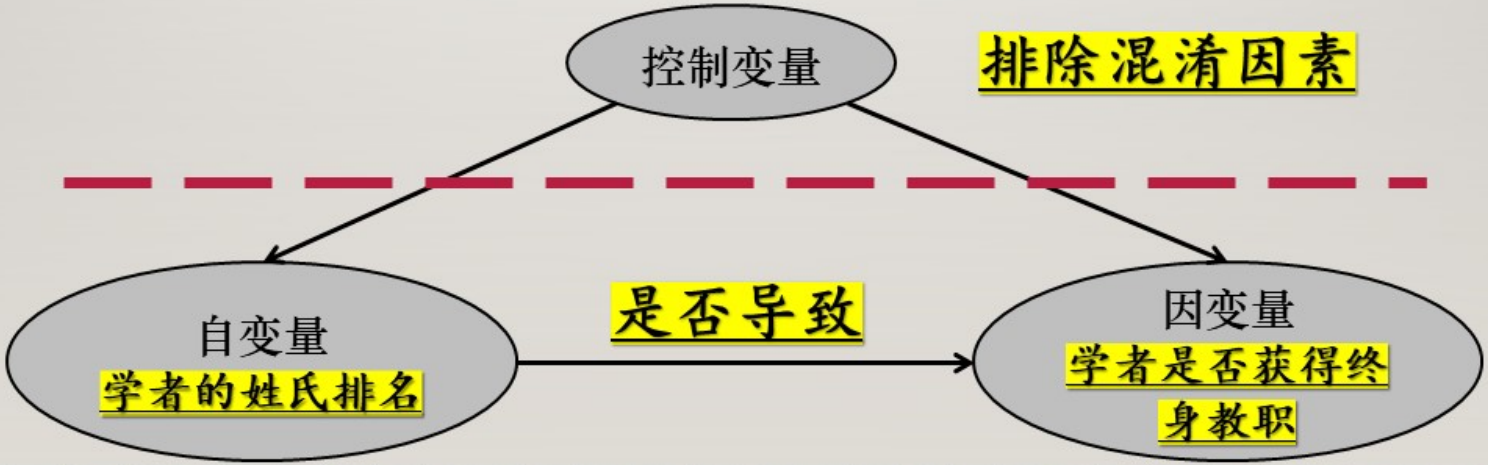
文本计量法的实施步骤：以姓氏排名和学术成就为例

- 概念操作化
- 控制变量（control variable）/混淆因素（confounding factors）
  - 学者的年龄、国籍、毕业院校、民族身份、宗教信仰……



文本计量法的实施步骤：以姓氏排名和学术成就为例

- 概念操作化
- 控制变量（control variable）/混淆因素（confounding factors）
  - 学者的年龄、国籍、毕业院校、民族身份、宗教信仰……



文本计量法的实施步骤：以姓氏排名和学术成就为例

- 将文本资料转化定量数据库
  - 美国35所大学经济学院的教师简历
  - 如Abhijit Banerjee

Curriculum Vitae
August 2019

ABHIJIT VINAYAK BANERJEE

DEPARTMENT: Economics

DATE OF BIRTH: February 21, 1961

CITIZENSHIP: US Citizen

EDUCATION

INSTITUTION	DEGREE	DATE
Harvard University Cambridge, Massachusetts	Ph.D.	1988
Jawaharlal University New Delhi, India	M.A.	1983
University of Calcutta Calcutta, India	B.Sc.	1981

TITLE OF DOCTORAL THESIS: *Essays in Information Economics*

自变量：
姓氏排名，编码为2

控制变量：
1、**年龄，编码为45**
（定距变量：取值0-100）

2、**国籍，编码为1**
（类别变量：美国=1，非美国=0）

3、**毕业院校，编码为1**
（类别变量：名牌大学=1，非名牌大学=0）

文本计量法的实施步骤：以姓氏排名和学术成就为例

- 将文本资料转化定量数据库
  - 如Abhijit Banerjee

PROFESSIONAL EXPERIENCE

ACADEMIC POSITIONS

2003- Ford Foundation International Professor of Economics, M.I.T.
2003- Director, Abdul Latif Jameel Poverty Action Lab, M.I.T.
1996-2003 Professor of Economics, M.I.T.
1994-1996 Associate Professor of Economics, M.I.T.
1993-1994 Pentti J.K. Kouri Career Development Associate
Professor of Economics, M.I.T.
1992-1993 Assistant Professor of Economics, Harvard University
1988-1992 Assistant Professor of Economics, Princeton University
1991 (Fall) Visiting Assistant Professor of Economics, Harvard University

FIELDS OF INTEREST

Economic Development
Information Theory
Theory of Income Distribution
Macroeconomics

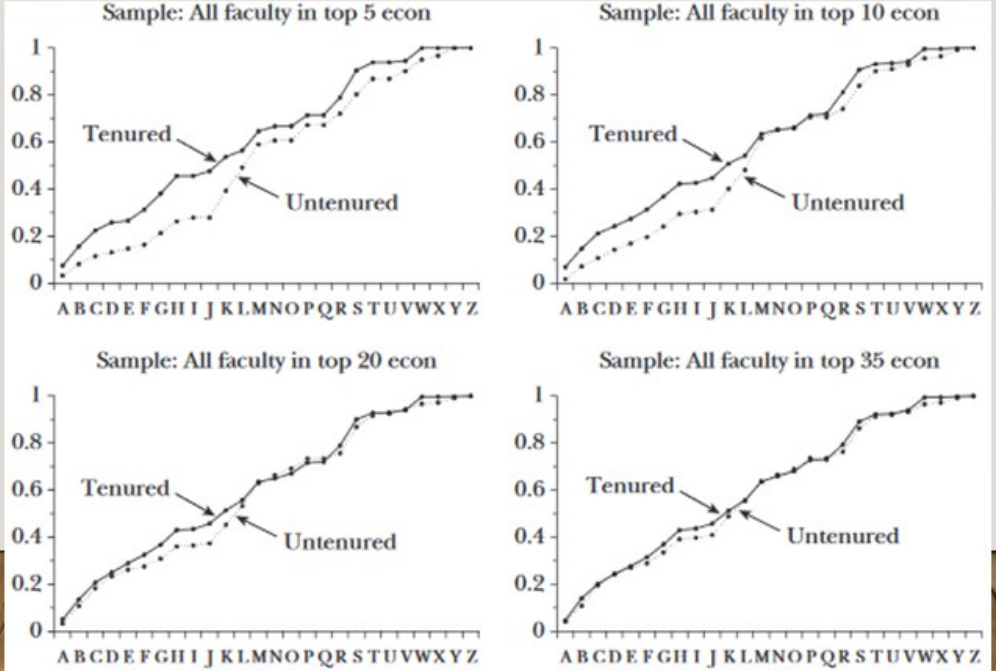
PROFESSIONAL SERVICES

Trustee, Save the Children, 2016 –
Member, United Nations High-level Panel on the Post-2015 Development Agenda, 2012 – 2013

因变量：
是否获得终身教职，编码为1
（类别变量：获得=1，未获得=0）

文本计量法的实施步骤：以姓氏排名和学术成就为例

- 选择适切统计技术进行分析
  - 描述统计：学者姓氏排名位置与是否获得终身教职的**累积分布图**



文本计量法的实施步骤：以姓氏排名和学术成就为例

- 选择适切统计技术进行分析
  - 因果关系分析：学者姓氏排名位置与是否获得终身教职的**线性概率回归分析**

Sample	*Top 5 econ*		*Top 10 econ*		*Top 20 econ*		*Top 35 econ*	
Last name initial	−0.0099**	−0.0086*	−0.0068**	−0.0063**	−0.0026	−0.0016	−0.0015	−0.0011
	(−2.18)	(−1.84)	(−2.08)	(−1.97)	(−1.12)	(−0.74)	(−0.84)	(−0.60)
American nationality	—	0.2282**	—	0.2062**	—	0.1873**	—	0.1436**
		(3.61)		(4.63)		(5.78)		(5.53)
Six origin controls	no	yes	no	yes	no	yes	no	yes
R^2	0.0225	0.1209	0.0106	0.1115	0.0016	0.0947	0.0006	0.0716
Number of obs.	208		405		799		1,233	
Number of tenured (%)	147 (70.7%)		293 (72.3%)		585 (73.2%)		911 (73.9%)	

拓展案例：学生姓名与学业成绩

- 骆明庆，《怡君比较会考试？名字、成绩与家庭背景》
- 在全台北高中学生中，名字越普通，大学联考成绩分数越低
  - 父母为子女取名不是随意的，而是反映父母的教育背景与经济地位

武亦姝　V.S.　沈佳宜

文本计量法的局限

- 许多文本资料的质量往往难以保证；
- 有些文本资料难以获得或无法获得；
- 原始文本资料缺乏标准化形式，难以编码和分析；
- 文本计量的效度和信度都存在一定的问题。