Explorations of cumulative advantage using data on French physicists

Bronwyn H. Hall and Jacques Mairesse

UC Berkeley, ENSAE-CREST, NBER, and MPI Munich

Two main aims

- Larger goal -- To investigate the performance of different structures of research systems within and cross country
- This paper -- identify the presence or absence of cumulative advantage in the presence of heterogeneous productivity
  - Some new results -- comparative behavior of the national laboratories vs the universities in French physics

Cumulative advantage

Morris Zapp in *Small World* (David Lodge, Penguin Edition, page 151), contemplates his impending visit to the Rockefeller Foundation at the Villa Serbelloni in Bellagio:

“The beauty of academic life is that to them that had had, more would be given…… All you needed to do to get started was to write one really damned good book, which admittedly wasn't easy when you were a young college teacher just beginning your career …… But on the strength of that one damned good book you could get a grant to write a second book in more favorable circumstances; with two books you got promotion, a lighter teaching load, and courses of your own devising; you could then use your teaching as a way of doing research for your next book, which you were thus able to produce all the more quickly. This productivity made you eligible for tenure, further promotion, more generous and prestigious research grants, more relief from routine teaching and administration. In theory, it was possible to wind up being full professor while doing nothing except to be permanently absent on some kind of sabbatical grant or fellowship.”

French research system - physics

- Two main institutions:
  - CNRS
    - Centralized selection, highly competitive at entry
    - After entry, essentially guaranteed salary for research until retirement
    - Can also teach but not required
  - Universities (centralized to some extent)
    - Similar early tenure, no research requirement although most do research, often in labs together with CNRS researchers
    - Little competition for academic staff across universities in different locations (unlike UK, US, etc)
  - Our data does not include PhD physicists at Grandes Ecoles or industry

Data

- Lists of the names, gender, and birth year for ~3300 French physicists (universe) collected in
  - 2005, 2015 – CNRS
  - 2005 – University
- When we compare the two, we use only the 2005 list for the CNRS
- Take to Open Alex, collect publications & citations up until 2022

	All	Sample	Balanced ages 21-55	Share female
CNRS 2005	994	966	842	15.9%
CNRS 2015 (not 2005)	496	473	46	19.6%
University 2005	1807	1600	1357	18.6%
Total	**3297**	**3039**	**2245**	**18.0%**

Selectivity

- Probability of being in sample
  - Increases slowly with birth year
  - For CNRS, does not depend on gender
  - For University physicists, marginal effect of gender is -16.4%
    - Suspect some problems with surnames, to be pursued.
- Exit probability (Weibull RE model)
  - Much lower for CNRS
  - Similar relationship to birthyear (slightly positive), gender (highly positive), and cumulative publications (highly negative)

Checking some stylized "facts"

1. The distribution of individual scientific productivity is extremely dispersed, as reflected by the fact that most papers, as well as the most highly cited papers, are published by a small share of scientists. (*Lotka 1926*)
    - In our data, 54% of pubs and 61% of cites generated by 20% of physicists (not as extreme as the usual formulation of Lotka's Law)
2. The resulting publication and citation hierarchy among scientists remains very stable during much of their lifetime (*David 1994*)
    - Rank correlations one year to next are 0.6-0.7, but correlogram declines to 0.4 by lag 11
3. The corresponding concentration tends to increase over time within scientific cohorts. (*Stephan 1996*)
    - Gini actually decreases over time rather than increasing.

Highly cited papers (cites>1000)

N of highly cited papers	Number of CNRS physicists	Number if random assignment	Number of Univ. physicists	Number if random assignment
Share with at least one	11.4%		3.1%	
0	1275	1188.5	1716	1700.4
1	91	208.7	29	58.4
2	39	33.6	14	2.1
3	19	6.6	1	.1
4	9	1.1	1	0
5	2	.4	0	0
6+	4	.1	0	0
Total	1439	1439.0	1761	1761.0

Chi-squared (7) = 267.4 (CNRS), 91.4 (Univ.) that assignment is random.

Heterogeneity or state dependence?

- Stylized facts about the distribution of scientific productivity only partly confirmed.
- Conditional on age quadratic:
  - Estimate individual productivities λ_i - about half reject Poisson
  - λ_is are approximately gamma distributed
  - => Negative binomial for pooled data
- Observed heterogeneity:
  - permanent differences in productivity
  - or does early success lead to more resources and time, and therefore increased productivity as time goes by?
- Answer: estimate FE models that include lagged cumulative productivity

Regression approach

- Model yearly paper and citation-weighted paper counts n_{it} as a function of age, birth year (cohort), year, gender, and the logarithm of cumulated past papers or cite-weighted papers.

$$E[n_{it} \mid \alpha_i, year, x's] = \lambda_{it} = \alpha_i + \gamma_t + \beta_1 age + \beta_2 age^2 + \beta_3 D(female) + \beta_4 \log\left(\sum_{s=1}^{t-1} n_{is}\right)$$

- Note:
  - birth year and gender are time-invariant and therefore not identified in FE model with α_i*s.*
  - Age, year, and cohort (FE) are collinear, and quadratic in age allows identification.

Estimation

- OLS with Log(n_{it}) as dependent variable
- Poisson with fixed effects
- Negative binomial with "fixed effects"
  - Not really FE on the means so time-invariant variables identified
- GMM on first-differenced equation with lagged x's as instruments:

$$E\left[n_{it} - n_{i,t+1} \frac{\lambda_{it}}{\lambda_{i,t+1}} \middle| x_{it} \right] = E\left[n_{it} - n_{i,t+1} \exp((x_{it} - x_{i,t+1})\beta \middle| x_{it} \right] = 0$$

  - (*Blundell et al; Crepon and Duguet*)

Cumulative advantage coefficient estimates

	OLS FE	Poisson FE	Neg. bin. “FE”	GMM FE	OLS FE	Poisson FE	Neg. bin. “FE”	GMM FE
Dep. Var.	Log pubs	Pubs	Pubs	Pubs	Log cites	Cites	Cites	Cites
CNRS physicists								
Log (cum. pubs	0.423	0.615	0.559	0.973	0.481	0.360	0.354	0.465
or cites) lagged	(0.008)	(0.019)	(0.008)	(0.091)	(0.009)	(0.021)	(0.004)	(0.068)
University physicists								
Log (cum. pubs	0.358	0.593	0.516	1.686	0.405	0.365	0.376	0.803
or cites) lagged	(0.008)	(0.024)	(0.009)	(0.357)	(0.009)	(0.024)	(0.004)	(0.252)
Other variables	Age, age squared, gender in NBFE, (year dummies)							

In practice, controlling for age, year dummies have little impact on the other estimates.

Comments on the Results

- Age quadratic is collinear with cumulative lagged publications, so the peak age tends to be very early. That is, cumulative pubs explain productivity better than age.
- Regardless of estimation model, lagged cumulative publications are highly predictive of current publications (as counts or cite-weighted counts).
- Compute fraction of variance explained by individual effects and fraction explained by cumulative pubs.
  - They are correlated (obviously), so depends on order in which computed.
  - Fixed effects clearly have stronger explanatory power

Components of the variance of pubs & cites

CNRS: 53,618 observaions on 1,435 scientists.					
	Pubs	Cites		Pubs	Cites
Age & age squared	21.6%	16.8%	Age & age squared	21.6%	16.8%
Add fixed effect	34.1%	26.1%	Add cumulative output	9.5%	8.5%
Add cumulative output	1.1%	1.5%	Add fixed effect	23.5%	17.8%
Unexplained	43.2%	55.6%	Unexplained	45.4%	56.9%

University: 60,634 observations on 1,565 scientists.					
	Pubs	Cites		Pubs	Cites
Age & age squared	18.1%	14.3%	Age & age squared	18.1%	14.3%
Add fixed effect	30.0%	25.1%	Add cumulative output	7.4%	7.1%
Add cumulative output	0.6%	0.6%	Add fixed effect	20.8%	16.7%
Unexplained	51.4%	60.0%	Unexplained	53.7%	62.0%

Reservations

- Highly selected group of physicists, especially CNRS.
- See Myers presentation – output measures may not capture everything that matters.
- Age is also highly correlated with cumulative pubs and cites, so captures some of their effect.
- And more……..