Coding Exercise

Kira Silvestrovich

Collecting data

I decided to analyze the connection between career tracks and publications activity of the Computer Science Faculty in the UC Santa Barbara.

Firstly, my main goal was to scrape as much data from the University website as possible to automate my process since it has a similar structure of pages for faculty members and does not ban requests.

So, my first step was using the BeautifulSoup package for web scrapping to collect data on Faculty names, links to their personal pages on the University Website links to personal websites, and links to sites with a list of publications.

In total, we work with 43 Faculty members (there were 63 people on the website, but 20 of them are affiliated faculty/ faculty emeriti). For all of them, we get names, links to personal pages and personal websites. For 20 of them, we get links to a list of publications.

I also tried to get their campus/off-campus affiliations from the website, however, the coverage is poor.

Then, I add LinkedIn pages links (I decided to do it manually as LinkedIn is currently one of the most difficult websites to scrape as they massively invested in defensive intelligence and quickly ban your api after scrapping even a few profiles).

Next, I looked at the links to the publication pages we received from the personal pages and realized that it will be better to have a similar structure of publications sites because f.e google scholar includes publications from conferences, but in a personal website, Professor may decide not to include them) and that is why I decided to get google scholars links for all Faculty for whom it is possible and for some of them we already have this link.

As the next step, I scraped the list of publications and years of these publications for each Faculty member who has a google scholar page. And then I counted the number of publications each year.

Afterwards, I imported dataset to excel and using Linkedin and CVs I add information on industry jobs (I do not consider an internship that is less than a year as an industry job).

Then, I import dataset to Stata. I prefer working with web scrapping/algorithms/interactive graphs in python while doing statistical analyzes in Stata. But of course, it is just my preference and if it will be needed to make all the work in one language, I will be able to do it.

Data Analyses

So, now having all information we want to investigate whether publication activity changes while Professors are engaged in work outside the academia.

In Stata, firstly I test the hypothesis that during the years while a person has an industry job, he/she has on average fewer publications than in years when he/she works only in academia. To test this hypothesis firstly, I make a binary variable having an industry job for those who at least once had one. Then, I use, ttest while controlling for unequal variances in the groups. We exclude 2022 from the analysis since it has just begun.

As a result, we see, that, on average, the number of publications in years when a person has an industry job is less in comparison with years when he does not (12.96 versus 15.7). However, this difference is insignificant.

But this result can be biased in several ways: firstly, the experience of the person is important, if we are talking about a professor who at the time of 2000 has 20 years of experience, he/she probably has a lot of RAs, which help carry out all the technical part of projects, probably he/she has tenure and there is no need to worry about the number of publications, if something goes wrong, in which case he/she can more easily combine work with work in the academy without compromising publications.

Secondly, there may have been years when number of publications declined in general for all Faculty. F.e. it can be 2020, when there was stress from switching to online.

Also, maybe the number of years in the industry affects, for example, if a person has been in the industry for one year, maybe that will dramatically affect the number of publications, and if a person has been combining academia and industry for ten years, the impact may not be as great for them.

Ideally, if we were to pursue this project further we would also want to control for position held, gender, age, the year when person started working in academia, marital status, etc.

At this point, going into the regression analysis, I would control for year, number of years in the industry, total number of publications as an approximation of work experience.

We get the result (table 1) that having an industry job, on average, leads to having 7 less publications a year and the result is significant.

Table 1: Number of publications and industry job

	(1)
having_industry_job	-7.374***
	(2.634)
total_publ	0.044***
	(0.007)
years_in_industry	0.391*
	(0.201)
Observations	263

Standard errors in parentheses

Dependent variable: number of publications. Year fixed effects are included

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Based on these results, our hypothesis that for professors with both academy and industry experience, the number of publications decreases while working in the industry is confirmed. However, we would need to add more control variables and extend the study to departments at other universities to continue the analysis and test for robustness.